FULCRUM: PAIR-DEPENDENT PREFERENCES BEYOND TRANSITIVE RANKINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many systems decide between two options by scoring each once and preferring the higher score. Whenever the score is a fixed function of the option alone, the resulting preference is transitive. Judgments on several criteria need not behave this way: the criterion that settles a comparison can change from pair to pair, and then no single scoring reproduces all the verdicts. We study FULCRUM, a comparison rule for structured argumentation in which the criterion that decides a pair is chosen from the pair itself (the criterion on which the two options differ most), and the resulting verdicts are treated as defeats between arguments and evaluated under standard acceptance semantics. We prove that this rule produces preference cycles of every length that no single ordering can represent, and we characterize, through pairwise disagreement and a minimum-weight feedback arc set, the maximum accuracy any single ordering can achieve on a given set of comparisons. We also show that the consistency guarantees of structured argumentation do not require transitivity: they hold under an explicit reasonableness condition on the ordering, and with conjunctive strict rules, the choice of how conjunctions are scored decides whether that condition holds. A soft-gated version of the rule is learnable from pairwise labels and attributes each verdict to a criterion. In a controlled setting whose true rule is gap-driven, it gains up to 7.3 accuracy points over a linear Bradley–Terry model and exceeds the single-ordering ceiling; on transitive preferences, it is neutral, and more flexible pairwise models do better when the dependence takes another form.

1 INTRODUCTION

A common way to decide between two options is to give each a score and prefer the higher score. Bradley–Terry reward models (Bradley & Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022) and rule-based reasoners with fixed rule priorities (Nute, 1994) both work this way. Whenever the score is a fixed function of the option alone, the deterministic preference it induces is a *single ordering*: each option has one position, the same in every comparison, so the preference is transitive on every triple (noisy observations of such a model can still violate transitivity). Multi-criteria judgments are different. Consider three reports scored on source reliability, specificity, and modal force, with $s(a) = (0, 1, 2)$, $s(b) = (1, 2, 0)$, $s(c) = (2, 0, 1)$. Deciding each pair on the criterion where the two reports differ most prefers a to b (modal force), b to c (specificity) and c to a (source reliability). Each pair disagrees most on a different criterion, so the deciding criterion rotates, and no assignment of one score per report reproduces the three verdicts (Figure 1). Resolving a conflict between two arguments is such a comparison, so the survivor relation can cycle in the same way.

Transitivity is thus a property of scalar scoring, not of multi-criteria judgment. Aggregating transitive criteria can yield an intransitive whole (May, 1954; Tversky, 1969; Fishburn, 1991); intransitive human comparisons have motivated alignment objectives built on general pairwise preference probabilities (Munos et al., 2024; Swamy et al., 2024; Wu et al., 2025); and matchup models represent cycles with latent vectors (Chen & Joachims, 2016; Balduzzi et al., 2018). Two questions then need precise answers: how much accuracy a single ordering must give up on a given set of comparisons, and whether giving up transitivity costs the consistency guarantees a reasoner needs. Formal argumentation answers both: preference turns attacks into defeats, Dung's semantics (Dung, 1995) decide acceptance, and rationality postulates (Caminada & Amgoud, 2007) define consistency.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

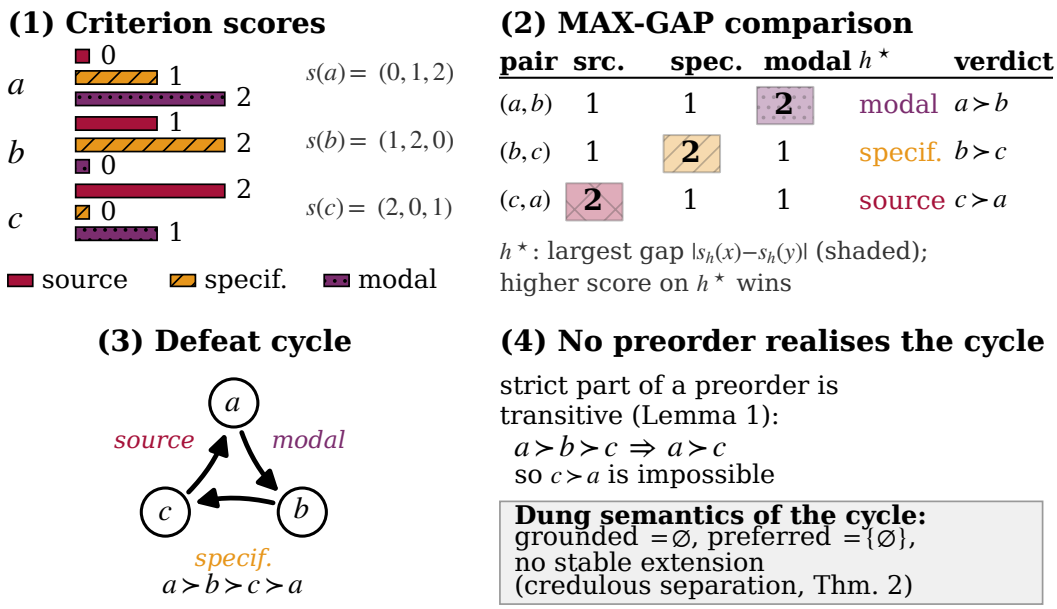


Figure 1: FULCRUM on a three-argument witness (color and hatch identify the criterion: crimson source, amber specificity, plum modal force). MAX-GAP selects the criterion with the largest pairwise score difference (shaded); the selected criterion rotates across pairs, producing $a \succ b \succ c \succ a$. No preorder can realize this cycle (Lemma 1); panel (4) reports its Dung semantics.

FULCRUM is the comparison rule of the example made formal: the criterion that decides a pair is chosen from the pair itself (MAX-GAP: the criterion of largest disagreement decides), and the verdicts are converted into defeats between arguments. The argumentation theory concerns this discrete rule; the learned comparator of Section 5 is a smooth relaxation of it, and only the comparator's pairwise predictions and attributions are evaluated experimentally. We make three contributions.

- *Representation.* MAX-GAP produces strict cycles of every length that no single ordering can represent; this separation holds for credulous acceptance. With one or two criteria, the rule can express exactly the linear orders, with n criteria, every orientation of n options, and for any fixed number of criteria, only a vanishing fraction of them; the last two bounds do not match, so the exact threshold is open. The *transitivity ceiling* gives the best accuracy any single ordering can reach on a fixed set of comparisons; its numerical value is certified or bracketed per dataset.
- *Rationality.* Applying the postulate theorems of Modgil & Prakken (2013), closure and consistency hold for any ordering that is reasonable in their sense, transitive or not. The cyclic ordering is reasonable when no strict rule has two or more antecedents; with conjunctive strict rules, the rule used to score a conjunction decides the outcome, since aggregations dominating the maximum always satisfy the required clause, while those bounded by the minimum fail on the constructed cycles. The results are conditional on well-definedness and on the stated aggregation.
- *Learning.* A soft-gated comparator with an antisymmetric score spans a fixed linear utility ($\beta = 0$) and the hard-gated rule ($\beta \to \infty$), decomposes its score by criterion, and satisfies an exact flip condition. On two controlled worlds and three empirical datasets, against transitive and cycle-capable baselines, it gains where the dependence is gap-driven, is neutral where preferences are transitive, and is outperformed by more flexible pairwise models where the dependence has another form.

Sections 2–6 develop these in turn.

2 Problem Setting and the Representational Gap

An argumentation framework (AF) is a pair $(\mathcal{A}, R)$: a set of arguments and an attack relation $R \subseteq \mathcal{A} \times \mathcal{A}$ (Dung, 1995). Acceptance is defined for sets of arguments. A set S is *conflict-free* if no two of its members attack each other, and S *defends* an argument a if every attacker of a

is itself attacked by some member of S. A conflict-free set that defends each of its members is *admissible*. Three standard extensions are used below: the *grounded* extension, the least fixed point of the map sending S to the arguments it defends (the most cautious position); *preferred* extensions, the inclusion-maximal admissible sets; and *stable* extensions, conflict-free sets that attack every outside argument. Grounded acceptance is skeptical; preferred and stable acceptance are credulous.

Preferences are entered by turning attacks into *defeats*: a defeats b if a attacks b and b is not strictly preferred to a, and the semantics are then applied to the defeat graph. Preference-based AFs (Amgoud & Cayrol, 1998) take the preference to be a preorder over arguments; value-based AFs (Bench-Capon, 2003) derive it from a total order over values; ASPIC$^+$ (Modgil & Prakken, 2013; 2014) derives an argument ordering from orderings on rules and premises; ABA$^+$ (Čyras & Toni, 2016) uses a preorder over assumptions. Amgoud & Vesic (2011) showed that removing attacks in this way can leave extensions that are not conflict-free when attacks are asymmetric, and applied preferences at the level of extensions instead. Extended AFs (Modgil, 2009) let arguments attack attacks, which can encode a hand-written cycle. In each construction reviewed in Appendix G and Table 14, the ordering that resolves attacks has an acyclic strict part, and none derives a cyclic preference from the arguments' own features.

Caminada & Amgoud (2007) ask that a structured formalism be closed under sub-arguments and strict rules and be directly and indirectly consistent. Modgil & Prakken (2013) prove these for ASPIC$^+$ when the framework is *well defined* (their Definition 12) and the argument ordering is *reasonable* (their Definition 18: an infallible argument is never below a fallible one, strict continuations of one argument keep its strength, the strict part is asymmetric, and a condition on strict continuations of *sets* holds). Transitivity is not among these conditions. A corrigendum (Modgil & Prakken, 2018) later revised how orderings are *derived* from premise and rule orderings (their Definitions 19–22); the postulate theorems and Definition 18 are unchanged (Appendix B).

On the learning side, Bradley–Terry reward models (Bradley & Terry, 1952; Ouyang et al., 2022; Rafailov et al., 2023) induce single orderings. Non-transitive preference learning represents cycles through latent matchup vectors (Chen & Joachims, 2016; Balduzzi et al., 2018) or through general pairwise preference probabilities inside an alignment objective (Munos et al., 2024; Swamy et al., 2024; Wu et al., 2025). FULCRUM differs in representation (a pair-dependent choice of criterion rather than free matchup vectors), in downstream use (defeat semantics with postulate guarantees) and in what a decision exposes (a contribution per criterion). We evaluate against the matchup-vector models and other pairwise comparators; the alignment objectives are not pairwise predictors and are not evaluated.

3 The FULCRUM Comparator and Defeat Construction

Fix a finite set $\mathcal{C}$ of m comparison criteria. Each argument $a \in \mathcal{A}$ carries a claim, a strength type $\tau(a) \in \{\text{CERTAIN}, \text{PROBABLE}\}$ (infallible or fallible), and a score vector $s(a) \in \mathbb{R}^{\mathcal{C}}$. For a pair write $d_h = s_h(a) - s_h(b)$.

A *criterion ordering* is a permutation of $\mathcal{C}$, read as a priority. A *policy* ρ assigns a criterion ordering to each unordered pair $\{a, b\}$; the static policy ρ_0 uses one fixed ordering for every pair (the lexicographic rule). A policy is *endogenous* if its choice for $\{a, b\}$ is determined by $s(a)$ and $s(b)$ alone; this matters for the separation, since a choice driven by an external tag could be absorbed by a single-ordering model into the argument's identity.

Definition 1 (MAX-GAP)**.** $\rho^\star(\{a, b\})$ orders the criteria by decreasing $|d_h|$, breaking ties by a fixed criterion order π.

Two PROBABLE arguments are compared by scanning $\rho(\{a, b\})$ in order: the first criterion on which their scores differ decides, and the higher score wins; if no criterion differs, the pair is *tied*. We call the result the *preponderance* relation $\succcurlyeq_\rho$: it records which argument carries more weight on the deciding criterion. Because ρ depends on the unordered pair, the same ordering is used whichever argument is listed first, so the strict part $\succ_\rho$ is asymmetric. A *strength gate* completes the rule: a CERTAIN argument is strictly preferred to a PROBABLE one without any scan, and two CERTAIN arguments are tied; Section 4.2 shows that this is exactly the fallibility clause that the postulate theorems require.

A mutual attack is resolved by the rule above (the strictly preferred argument's attack survives and the reverse attack is dropped; a tied pair keeps both); a one-way attack is dropped only when its target is strictly preferred. Grounded, preferred and stable extensions are computed on the defeat graph. Appendix C proves that the map is well defined, never dissolves a mutual conflict, keeps every extension conflict-free even when $\succcurlyeq_\rho$ is intransitive, and costs $O(|R|\, m)$, so acceptance keeps Dung's complexity (Dvořák & Dunne, 2017).

4 Representational and Rationality Results

4.1 Separation and the size of the gap

Lemma 1 (Transitivity barrier). *Let $\succcurlyeq$ be a preorder on $\mathcal{A}$ and $\succ$ its strict part. Then $\succ$ is irreflexive and transitive, hence acyclic.*

Each construction of Section 2 supplies a preference whose strict part is acyclic (for the ASPIC$^+$ liftings see Modgil & Prakken, 2013, Props. 23–24); a Bradley–Terry utility induces a weak order, a linear order under any globally consistent tie-breaking. On the witness $a = (0, 1, 2)$, $b = (1, 2, 0)$, $c = (2, 0, 1)$ the gaps are $(1, 1, 2)$, $(1, 2, 1)$, $(2, 1, 1)$ for $(a, b), (b, c), (c, a)$, each with a unique maximum, so MAX-GAP yields $a \succ b \succ c \succ a$.

Theorem 1 (Comparator separation). *No preorder on $\{a, b, c\}$ has a strict part equal to the $\rho^\star$-preponderance relation.*

Theorem 2 (Extension separation, credulous semantics). *With R the three mutual attacks on $\{a, b, c\}$, the $\rho^\star$ defeat relation is the directed 3-cycle, with grounded $= \emptyset$, preferred $= \{\emptyset\}$ and no stable extension; no preorder induces this defeat relation nor its (preferred, stable) profile.*

Theorem 3 (Cycles of every order). *For $k \geq 3$ let $v_s[i] = (i + s) \bmod k$ over k criteria. Under* MAX-GAP *the preponderance relation contains the strict k-cycle $v_0 \succ v_1 \succ \cdots \succ v_{k-1} \succ v_0$ (for $k > 3$ the relation also orients the non-consecutive pairs).*

Proofs are in Appendix A. The separation is credulous: a directed 3-cycle and a symmetric 3-clique both have grounded extension $\emptyset$. The cycle is not tuned: the grid $\{0, \ldots, 3\}^3$ contains 754 such cycles, 68 of them with a unique largest gap on every pair (Appendix D).

How much can a pair-dependent rule represent, and how much must a single ordering forgo? Say MAX-GAP with m criteria *robustly realizes* a tournament T on n arguments if there are scores such that every pair has a unique largest gap, and the verdicts orient every pair as T does.

Theorem 4 (Capacity of MAX-GAP). *(a) With $m \leq 2$ criteria the robustly realisable tournaments are exactly the $n!$ linear orders: for $m = 2$ the winner always has the strictly larger coordinate sum. (b) With $m \geq 3$ intransitive tournaments are realized, and with $m = n$* every *tournament on n arguments is robustly realised, with scores in $\{0, 1, 3\}$. (c) For every m the number of robustly realisable tournaments is at most $\big(2e(n-1)(m+1)\big)^{nm}$; hence for fixed m the realisable fraction of the $2^{\binom{n}{2}}$ tournaments tends to 0, and universality requires $m \geq (n-1)/\big(2\log_2(2e(n-1)(m+1))\big) = \Omega(n/\log n)$.*

Part (c) uses Warren's bound on sign patterns of polynomials (Warren, 1968). The theorem locates MAX-GAP between two extremes: with one or two criteria, it is a single global order, and with as many criteria as options, it expresses any orientation. Part (c) shows how the number of criteria controls reach: for fixed m the reachable fraction vanishes as n grows, and reaching all orientations needs $\Omega(n/\log n)$ criteria; the sufficient $m = n$ and this necessary bound do not meet, so the exact threshold is open.

Theorem 5 (Transitivity ceiling). *Let D be a finite multiset of N pairwise outcomes over items V, with n_{ij} the number of times i beat j. Let $L_{\text{pair}} = \sum_{i<j} \min(n_{ij}, n_{ji})$ and let T be the majority digraph with arc weights $|n_{ij} - n_{ji}|$. The maximum number of outcomes in D that any linear order on V predicts correctly is $N - L_{\text{pair}} - \text{FAS}_w(T)$, where $\text{FAS}_w(T)$ is the minimum weight of a feedback arc set of T. Consequently, every model whose verdicts are induced by a linear order has accuracy at most $1 - (L_{\text{pair}} + \text{FAS}_w(T))/N$, while a pairwise predictor is bounded only by $1 - L_{\text{pair}}/N$.*

Corollary 1 (Triangle bound). *For a complete deterministic tournament on n items with a fraction f_3 of cyclic triples, every linear order has accuracy at most $1 - f_3/3$.*

The ceiling is the Slater/Kemeny optimum (Slater, 1961; Kemeny, 1959), NP-hard on tournaments (Alon, 2006; Charbit et al., 2007) but solvable at our sizes; we solve it by a lazily constrained linear-ordering integer program, exactly where the solver certifies optimality and as a certified bracket otherwise (Appendix F). The triangle bound is closed-form but loose (Section 6). On a fixed comparison set the theorem separates the irreducible *representational gap* $1 - L_{\text{pair}}/N$ − ceiling, which no single ordering can close, from the gap ceiling − accuracy between the unrestricted hindsight optimum and a particular transitive model, a gap that can include feature restrictions, approximation, optimization and finite-sample effects. On a matching, where each item appears in one comparison, the ceiling is 1 and uninformative.

4.2 Rationality postulates without transitivity

Corollary 2 (Application of the ASPIC$^+$ postulate theorems)**.** *Let a structured* FULCRUM *framework be well defined (Modgil & Prakken, 2013, Def. 12) and let its argument ordering, with strict part $\succ$ given by Section 3, be reasonable (Modgil & Prakken, 2013, Def. 18). Then by their Theorems 12–15 every attack-complete extension is closed under sub-arguments and strict rules and is indirectly consistent, and every attack-admissible extension is directly consistent, whether or not $\succ$ is transitive.*

Appendix B lists every hypothesis and maps FULCRUM objects to the source definitions: CERTAIN arguments are *encoded* as strict-and-firm arguments (axiom premises and strict rules only), PROBABLE arguments contain an ordinary premise or a defeasible rule, a fallible argument's preponderance score is the score of its top defeasible rule or ordinary premise, and a strict continuation of one argument inherits that score unchanged (clause 18-1(iii)). The corollary is an application, not a re-proof; its content is that transitivity is absent from the hypotheses. What remains is the set clause of Definition 18-2: for $\{C_1, \ldots, C_n\}$ it must not be that every C_i strictly beats a strict continuation C_i' of the other members. By Definition 17 of the source, a strict continuation of a set contains *exactly* the union of the members' ordinary premises and defeasible rules.

Theorem 6 (Sufficient condition, and failure under combining rules)**.** *Assume* CERTAIN *arguments are encoded as strict-and-firm and scores pass unchanged through strict continuations of single arguments. (a) If no strict rule has two or more antecedents, then no strict continuation of a set of two or more independent* PROBABLE *arguments (disjoint, nonempty fallible material) exists, clause 18-2 holds vacuously, the ordering is reasonable, and Corollary 2 yields all four postulates for the intransitive cycle. (b) By combining strict rules, canonically conjunction introduction $A, B \Rightarrow A \wedge B$, there are orderings for which clause 18-2 fails.*

Part (b) depends on how a conjunctive continuation is *scored* for preponderance. Let a symmetric coordinatewise aggregation g assign to a strict continuation of $\{C_j\}_{j\in S}$ the score vector $g(\{s(C_j)\}_{j\in S})$, with $g(\{s\}) = s$. Call g *max-dominating* if $g(S)_h \geq \max_{j\in S} s_{j,h}$ for all S, h (sum of nonnegative scores; maximum) and *min-dominated* if $0 \leq g(S)_h \leq \min_{j\in S} s_{j,h}$ (weakest link).

Theorem 7 (Aggregation decides the set clause)**.** *Under* MAX-GAP *with any fixed tie-break: (a) if g is max-dominating, the set clause of Definition 18-2 holds for* every *finite set of fallible arguments and every choice of strict continuations; (b) if g is min-dominated, the k-cycle witness with conjunction introduction violates the clause for every $k \geq 3$.*

The proof of (a) shows that if every C_i beat the aggregate of the others, each would be the unique maximum on its own deciding criterion, forcing a tie in gaps between every two such criteria that a fixed tie-break cannot resolve in both directions. Sum and max satisfy (a); weakest link, the ASPIC$^+$ default for logical strength, satisfies (b); the mean satisfies neither (it holds on 66 of the 68 unique-largest-gap cyclic triples of the $\{0, \ldots, 3\}^3$ grid, on which weakest link holds for 36, and fails on 153 of 20,000 random sets; Appendix D).

The theorem identifies the one additional construction choice, how a conjunction is scored for preponderance, on which preservation of the postulates turns once combining rules are present. One may keep weakest link for logical strength and a max-dominating aggregation for preponderance; a proof that this combined ordering satisfies all of Definition 18 in a specific instantiation is not given here and remains open. The postulate theorems themselves are not new; what is new is the intransitive ordering and the explicit conditions under which it is reasonable.

5 A LEARNABLE COMPARATOR WITH CRITERION-LEVEL ATTRIBUTION

The discrete scan has a learnable counterpart, in the spirit of quantitative argumentation (Potyka, 2021; Baroni et al., 2018). Write $d = s(a) - s(b) \in \mathbb{R}^m$ for the score differences. Three pieces are kept separate: the mechanism that selects the deciding criterion, the criterion-specific contributions, and the comparison score, which is antisymmetric.

With a *hard gate*, the selected criterion is $c(a,b) = \arg\max_h |d_h|$, the criterion of largest disagreement, and the score is $z = W_c \cdot d$, where $W \in \mathbb{R}^{m \times m}$ holds one weight vector per selectable criterion. With $W = I$, no strength term and a unique largest gap, the sign of z is the MAX-GAP verdict; a general W, a strength term or a tied largest gap can change it. A *soft gate* replaces the argmax by a distribution over criteria:

$$g_c(a,b) = \mathrm{softmax}_c\big(\beta |d_c| + b_c\big), \qquad \text{(criterion weights)} \tag{1}$$

$$z(a,b) = \textstyle\sum_c g_c(a,b)\, W_c \cdot d + \lambda(\sigma_a - \sigma_b), \qquad \text{(comparison score)} \tag{2}$$

$$P(a \succ b) = \sigma\big(z(a,b)/T\big), \qquad \text{(probability)} \tag{3}$$

where $\beta > 0$ is a learned sharpness, $b \in \mathbb{R}^m$ a learned bias over criteria, $\sigma_a, \sigma_b \in [0,1]$ scalar strengths (1 for CERTAIN), $\lambda \geq 0$ a learned strength gate, $\sigma(\cdot)$ the logistic function and T a temperature fitted on validation data after training (Guo et al., 2017). The criterion weights (1) depend on $|d|$ and are therefore the same whichever argument is listed first; since $d(b,a) = -d(a,b)$, the score (2) satisfies $z(b,a) = -z(a,b)$. Writing $\bar{w}_h = \sum_c g_c W_{c,h}$, the score decomposes by criterion into *criterion contributions* $\kappa_h = \bar{w}_h d_h$ with $\sum_h \kappa_h = z - \lambda(\sigma_a - \sigma_b)$, and the attributed criterion is $h^\star = \arg\max_h |\kappa_h|$. Two limits matter: at $\beta = 0$ the weights are constant, so z is a single linear utility difference, and the induced preference is transitive; as $\beta \to \infty$ the weights concentrate on the largest gap and the model becomes the hard-gated comparator on every pair whose largest gap is unique.

The predictor evaluated in Section 6 is $z(a,b) + b_0$: the antisymmetric score plus one scalar *position intercept* b_0, shared by all models, which absorbs any effect of which option is listed first. The antisymmetry and transitivity statements concern z: a model is called transitive when $z = u(a) - u(b)$ is a difference of utilities, so that its deterministic preference is a single ordering; adding b_0 shifts the decision threshold, and with $b_0 \neq 0$ the full predictor's verdicts are no longer exactly induced by that ordering (Appendix F states where this affects a reported metric).

Proposition 1 (Flip conditions)**.** *(a) For the discrete comparator with policy ρ and a non-equipollent* PROBABLE–PROBABLE *pair (a verdict governed by the criterion scan, not by the strength gate) with deciding criterion $h^\star$, swapping the two arguments' scores on $h^\star$ reverses the verdict* if *$h^\star$ is again the first discriminating criterion of ρ on the swapped pair; otherwise the swapped verdict is the sign of d on whichever criterion then comes first, and may still reverse (Appendix A). (b) This holds for every pair if ρ is* swap-invariant*, i.e. unchanged when $s_h(a)$ and $s_h(b)$ are exchanged on any single criterion; static, per-tag and* MAX-GAP *policies (any policy depending on the pair only through $|d|$) are swap-invariant. (c) The assumption is needed: a symmetric endogenous policy that is not swap-invariant can fail to flip (Appendix A). (d) For the score z of the hard- or soft-gated model, with strengths and all non-intervened terms held fixed, the score swap on $h^\star$ maps $z \mapsto z - 2\kappa_{h^\star}$, so the strict sign of the verdict reverses iff $z(z - 2\kappa_{h^\star}) < 0$, equivalently $|\kappa_{h^\star}| > |z - \kappa_{h^\star}|$; equality gives a zero logit, a tie rather than a reversed verdict.*

Part (d) makes the flip rate a measurable property of a fitted model rather than a guarantee; a large contribution is not by itself evidence that the attributed criterion drives the verdict, which Section 6 tests by intervention.

6 EXPERIMENTS

We ask three questions: how much a pair-dependent comparator gains over a single ordering when criteria conflict, relative to the ceiling of Theorem 5 and to other cycle-capable models (Section 6.1); whether the full comparator helps on empirical multi-criteria data, and where (Section 6.2); and whether the attributed criterion carries the verdict under intervention (Section 6.3).

Protocol. Every model is evaluated on ten seeds; a seed regenerates a controlled world or resplits an empirical dataset, and all models of a seed share the same splits. We report mean held-out accuracy

Table 1: Held-out pairwise accuracy (mean $\pm$ t-based 95% half-width over 10 seeds) on the two controlled worlds and three empirical datasets; protocol in Section 6. T: transitive model. Bold marks the best learned non-oracle model in each column. Ceilings (Theorem 5): exact on every controlled test tournament (10 seeds); on Pokémon and Sushi-B computed on seeds 0–2 only (a certified bracket, and a value bounded within 0.001); trivially 1 on the HelpSteer2 matching. Full tables with NLL, Brier, ECE, paired gaps and every baseline: Appendix F.

model	World C, r=0	World M	Pokémon	Sushi-B	HelpSteer2-Pref.
BT (linear utility)$^{\mathrm{T}}$	0.815 $\pm$0.010	0.879 $\pm$0.024	0.887 $\pm$0.002	0.656 $\pm$0.003	0.921 $\pm$0.002
BT (MLP utility)$^{\mathrm{T}}$	0.806 $\pm$0.011	0.864 $\pm$0.024	0.885 $\pm$0.003	0.659 $\pm$0.003	**0.929** $\pm$0.004
lexicographic$^{\mathrm{T}}$	0.690 $\pm$0.025	0.686 $\pm$0.029	0.926 $\pm$0.002	0.641 $\pm$0.003	0.920 $\pm$0.000
mixture of utilities	0.820 $\pm$0.011	0.881 $\pm$0.025	0.934 $\pm$0.002	0.656 $\pm$0.003	0.919 $\pm$0.001
antisym. bilinear	0.812 $\pm$0.009	0.869 $\pm$0.027	0.879 $\pm$0.002	0.656 $\pm$0.003	0.925 $\pm$0.004
pair-MLP (matched)	0.818 $\pm$0.016	0.879 $\pm$0.026	0.939 $\pm$0.002	0.654 $\pm$0.004	0.925 $\pm$0.002
pair-MLP (large)	**0.972** $\pm$0.002	0.914 $\pm$0.012	**0.946** $\pm$0.002	0.658 $\pm$0.004	0.928 $\pm$0.004
blade-chest (item emb.)	–	–	0.929 $\pm$0.002	**0.661** $\pm$0.003	–
FULCRUM-hard	0.862 $\pm$0.009	0.915 $\pm$0.012	0.897 $\pm$0.002	0.656 $\pm$0.003	0.921 $\pm$0.002
FULCRUM-soft	0.875 $\pm$0.009	**0.952** $\pm$0.008	0.906 $\pm$0.002	0.655 $\pm$0.003	0.928 $\pm$0.002
transitive ceiling	0.883 $\pm$0.009	0.931 $\pm$0.016	[0.938, 0.952]	0.672	1.000
status (seeds)	exact (10)	exact (10)	bounded (0–2)	bounded (0–2)	trivial
pair ceiling	1.000	1.000	0.999	0.691	1.000

with a t-based 95% half-width, and gaps between models as paired per-seed differences in percentage points with a paired t-based 95% half-width. Ceilings are computed exactly, with a certificate of optimality, on every controlled test tournament; on Pokémon and Sushi-B the solver was run on seeds 0–2 only (a certified bracket on Pokémon; two exact values and one bounded within 0.001 on Sushi-B); on HelpSteer2, where each response appears once, the ceiling is trivially 1. Gaps to a ceiling use shared seeds and splits only (Appendix F).

Data. *World C* (majority-generated comparisons): 60 items with $m = 5$ criteria correlated at level r; item i beats j when it is higher on a majority of criteria; train, validation and test items are disjoint, and $r = 1$ is a transitive control. *World M* (gap-generated comparisons): $m = 5$ independent criteria with hidden per-criterion scales, and i beats j by MAX-GAP on the rescaled scores; the labels used for training and validation carry 10% noise, and test labels are clean. *Pokémon*: 50,000 battle outcomes among 784 Pokémon from a public dataset whose outcomes were generated by a simplified battle algorithm rather than recorded from play (Kaggle, 2017); the seven criteria are the six base-stat differences and a pair-level type-advantage term. *Sushi-B*: 5,000 respondents' rankings of 10 of 100 sushi items (225,000 comparisons) with six item attributes as criteria; its population majority has 1.1% cyclic triples (Kamishima, 2003). *HelpSteer2-Preference*: human pairwise preferences between responses with five annotated aspects as criteria; each response appears once (Wang et al., 2024a).

Models. Three families are compared. *Transitive models*, whose score is a utility difference: Bradley–Terry with a linear utility (BT), BT with a two-hidden-layer utility (BT-MLP), and a fixed lexicographic criterion order. *Generic cycle-capable models*: a mixture of linear utilities, an antisymmetric low-rank bilinear model, the blade-chest model with item embeddings (Chen & Joachims, 2016) where items recur, and an antisymmetrised pair-MLP at two sizes, one parameter-matched to FULCRUM and one large. *Gated comparators*: FULCRUM-hard and FULCRUM-soft, with the fixed MAX-GAP rule, a Condorcet rule, a linear model in d, and random- and sum-gated variants as ablations (Appendix F); the linear model in d is transitive except on Pokémon, where d contains the pair-level term. Trainable models minimise cross-entropy with early stopping on validation loss; the two fixed rules fit only polarities and a scale; the lexicographic order is chosen by search; every model is temperature-scaled on validation data and evaluated with the position intercept of Section 5.

6.1 CONFLICT, CEILINGS AND CYCLE-CAPABLE BASELINES

In World C the pair-dependent comparator gains over a single ordering as the criteria decorrelate. At $r = 1$ every model is near perfect (0.99–1.00). At $r = 0$ BT reaches 0.815 against a ceiling of 0.883, so 0.117 of its error is unreachable by any single ordering.

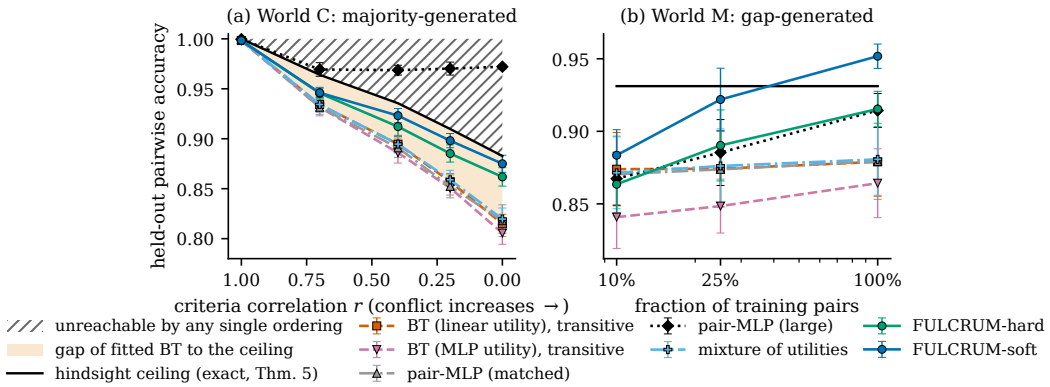


Figure 2: (a) World C: held-out accuracy as the criteria decorrelate. The black line is the exact ceiling of Theorem 5, the accuracy of the best single ordering chosen in hindsight on each test tournament; the hatched region above it cannot be reached by any single ordering, and the shaded band is the gap between the fitted linear BT model and that ceiling. (b) World M against the fraction of training pairs; the ceiling is a mean over the ten test tournaments. Markers: mean $\pm$ t-based 95% half-width over ten seeds; test labels are noise-free.

BT itself falls 0.068 short of that ceiling. FULCRUM-soft gains +6.0 $\pm$ 0.6 percentage points over BT and FULCRUM-hard +4.7 $\pm$ 0.4; the transitive BT-MLP does not improve over BT (-0.9 $\pm$ 0.7), so a nonlinear utility alone does not explain the gain.

Against the generic cycle-capable models the picture is two-sided. At matched parameter count the pair-MLP (+0.3 $\pm$ 1.0) and the mixture of utilities (+0.5 $\pm$ 0.6) do not gain, whereas the large pair-MLP gains +15.7 $\pm$ 1.0 and comes close to the majority rule; the fitted gated comparators stay below the ceiling at every r and only the large pair-MLP enters the region above it (Figure 2(a); parameter counts in Figure 5, appendix). Figure 4 (appendix) and Table 11 show where the gain lives: models are nearly indistinguishable on pairs where all criteria agree, and the gap widens as more criteria dissent, in both worlds and on Pokémon. On pairs that lie on a cyclic triple of the oracle, FULCRUM-soft reaches 0.827 against 0.741 for BT, and it reproduces the orientation of 0.20 of the cyclic test triples, a quantity that is 0 for a single ordering (Lemma 1) and 0.00 for the transitive models here.

World M (Figure 2(b), Table 6) is the case the gated comparator is designed for: FULCRUM-soft reaches 0.952 against 0.879 for BT, above the ceiling of 0.931 and above the large pair-MLP (0.914) with 32 parameters against 1,090 (fitted gate in Figure 6, appendix). With 25% of the training pairs (442 comparisons) the soft gate gains +4.8 $\pm$ 3.3 points over BT; with 10% (177) its gain is +1.0 $\pm$ 2.9, an interval that includes zero, and the hard gate (-1.0 $\pm$ 3.5) and large pair-MLP (-0.7 $\pm$ 1.2) do not exceed BT; smaller fractions were not tested.

Ablations (Table 7) show that a random or sum-based selection removes the gain and that the fixed-weight MAX-GAP rule falls below BT, so on these data both the gap-based selection and the learned per-criterion weights are needed. Shuffling the identity of the criteria per pair leaves World C unchanged, whose majority rule is symmetric in the criteria, but in World M reduces FULCRUM-soft's gain from +7.3 to +2.9 points, the level of the fixed MAX-GAP rule, and removes the large pair-MLP's gain (+1.1 $\pm$ 2.0). Within this experiment, gap-based selection and criterion identity contribute separately; this does not establish them as the only sources of the gain in general. Robustness to label noise and training-set size in World C is in Appendix F.

6.2 Empirical datasets

On the empirical datasets the full comparator gives a modest gain at best, and the dependence in the data is not primarily of the largest-gap kind (every model's paired gain is plotted in Figure 3, appendix). On Pokémon (Table 1; slices in Table 10) FULCRUM-soft improves on BT by +1.9 $\pm$ 0.1 points, with a gain of similar size across the prespecified closeness slices. A speed-first lexicographic order (0.926), the mixture (0.934), the matched pair-MLP (0.939) and blade-chest (0.929) all exceed FULCRUM-soft (0.906), and the large pair-MLP reaches 0.946. The ceiling bracket $[0.938, 0.952]$ (seeds 0–2, same splits) lies above FULCRUM-soft but contains the large pair-MLP's value, so it does

not settle whether that model exceeds the best single ordering, and the experiment does not separate feature restriction, approximation, optimisation and finite-sample effects in FULCRUM-soft's shortfall.

On Sushi-B the best single ordering in hindsight reaches 0.672 against 0.656 for BT, and the ceiling for an unrestricted pairwise predictor is 0.691 (seeds 0–2), so intransitivity can contribute at most about two points on this population preference. The largest paired gains over BT are small: $+0.4 \pm 0.1$ points for blade-chest with item embeddings and $+0.3 \pm 0.1$ for BT-MLP, while FULCRUM-soft's paired interval includes zero. On HelpSteer2-Preference the aspect-gated comparator improves on the linear aspect aggregate by +0.7 $\pm$ 0.3 points, and the transitive BT-MLP gains +0.8 $\pm$ 0.5; the improvement is consistent with a nonlinear map from aspects to preference and is not evidence of intransitivity, which a matching cannot exhibit whether or not the underlying population has it.

6.3 Attribution under intervention

The attributed criterion carries the verdict far more often than a control criterion does. Table 12 intervenes on $h^\star$: for FULCRUM-soft in World C, reversing the sign of $d_{h^\star}$ flips 56% of verdicts against 26% for a random other criterion, and keeping only $h^\star$ reproduces 78% of verdicts; the hard gate behaves alike, and the linear model in d scores comparably on the same diagnostics. By Proposition 1(d) the flip rate equals the fraction of pairs on which the attributed contribution exceeds the net of the other terms (Figure 7, appendix). These measure how the fitted models respond to these particular perturbations; they do not establish that the attribution explains the model's behaviour in general.

Discussion. Intransitivity determines the room above a single ordering, as quantified by the ceiling; the gain a learned comparator realises depends additionally on its inductive bias, data and optimisation. World M, whose labels are gap-generated, gives the strongest positive evidence for the gap-gated comparator; in World C cycle-capable models close part of the room above the ceiling and a sufficiently flexible pair-MLP does best; on Pokémon FULCRUM improves modestly over BT but the dependence is not primarily of the largest-gap kind; on Sushi-B little is available to any model; on HelpSteer2 the improvement is consistent with nonlinearity. What the gated comparator adds is structure: a contribution per criterion with a measured flip condition, and the defeat semantics and postulate conditions of Sections 3–4, which these pairwise benchmarks do not test.

7 Conclusion

Limitations. Theoretically, the extension-level separation is credulous; the postulate results are conditional on well-definedness and on the aggregation used for preponderance, with no reasonableness proof for a combined ordering in a specific instantiation; the capacity theorem concerns robust realizations, not learnability from finite data. The ceiling is a hindsight quantity on the given comparison multiset; it is loose on sparse graphs, on Pokémon the certified bracket does not resolve whether the best generic comparator exceeds the optimum ordering, and on HelpSteer2's matching it is trivially 1, so the observed graph is silent about cycles in the population. Empirically, the gap-based inductive bias pays only where the dependence is gap-driven (generic nonlinear comparators fit the majority rule and the Pokémon data better, and on Sushi-B, no model gains more than half a point over BT); the reported cycle-orientation values are those of the evaluated predictor, intercept included, not of the underlying ordering (Appendix F); the criterion scores await domain-expert validation, and the natural-language benchmark we run has fixed priorities (Appendix E).

Conclusion. A pair-dependent choice of the deciding criterion produces intransitive defeat that no single ordering can carry; on a given comparison multiset, the price of transitivity has an exact hindsight characterization as a feedback-arc-set fraction; and the postulates require a reasonable ordering, not a transitive one, with the scoring of conjunctions as the operative condition. A soft-gated realization makes the structure learnable and attributable, at the cost of an inductive bias that pays only in some settings.

Reproducibility statement

Complete proofs of Lemma 1, Theorems 1–7, Corollary 1 and Proposition 1 are in Appendix A; the hypotheses and object mapping for Corollary 2 are in Appendix B; the defeat-map properties

are proved in Appendix C. Appendix F specifies every dataset, split, seed, model, parameter count, optimization and stopping rule, calibration step, metric, interval procedure, and the ceiling solver. The finite computations reported alongside the theorems (the 29 and 355 preorders, the 68 grid cycles, the k-cycles for $k \leq 7$, the capacity construction for $n \leq 5$, the aggregation theorem on $20{,}000$ random sets, and the flip counterexample) are exhaustive checks of finite cases, not proofs of the general statements. Code and stored results sufficient to reproduce every number, table, and figure are provided as anonymized supplementary material.

AI USE STATEMENT

Generative AI tools assisted with language editing and polishing the manuscript. The authors reviewed the final text and take full responsibility for all scientific claims, analyses, and conclusions.

REFERENCES

Noga Alon. Ranking tournaments. *SIAM Journal on Discrete Mathematics*, 20(1):137–142, 2006. doi: 10.1137/050623905.

Leila Amgoud and Jonathan Ben-Naim. Ranking-based semantics for argumentation frameworks. In *Scalable Uncertainty Management (SUM)*, volume 8078 of *Lecture Notes in Computer Science*, pp. 134–147, 2013. doi: 10.1007/978-3-642-40381-1_11.

Leila Amgoud and Claudette Cayrol. On the acceptability of arguments in preference-based argumentation. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 1–7, 1998.

Leila Amgoud and Srdjan Vesic. A new approach for preference-based argumentation frameworks. *Annals of Mathematics and Artificial Intelligence*, 63(2):149–183, 2011. doi: 10.1007/s10472-011-9271-9.

David Balduzzi, Karl Tuyls, Julien Pérolat, and Thore Graepel. Re-evaluating evaluation. In *Advances in Neural Information Processing Systems*, volume 31, 2018. arXiv:1806.02643.

Pietro Baroni, Antonio Rago, and Francesca Toni. How many properties do we need for gradual argumentation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11544.

Trevor J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003. doi: 10.1093/logcom/13.3.429.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5–6):286–310, 2007. doi: 10.1016/j.artint.2007.02.003.

Pierre Charbit, Stéphan Thomassé, and Anders Yeo. The minimum feedback arc set problem is NP-hard for tournaments. *Combinatorics, Probability and Computing*, 16(1):1–4, 2007. doi: 10.1017/S0963548306007887.

Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 227–236, 2016. doi: 10.1145/2835776.2835787.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kristijonas Čyras and Francesca Toni. ABA$^+$: Assumption-based argumentation with preferences. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pp. 553–556, 2016.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995. doi: 10.1016/0004-3702(94)00041-X.

Wolfgang Dvořák and Paul E. Dunne. Computational problems in formal argumentation and their complexity. *IfCoLog Journal of Logics and their Applications*, 4(8), 2017. Reprinted in the Handbook of Formal Argumentation, College Publications, 2018.

Xiaotong Fang, Zhaoqun Li, Chen Chen, and Beishui Liao. LLM-ASPIC$^+$: A neuro-symbolic framework for defeasible reasoning. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 1567–1574, 2025. doi: 10.3233/FAIA250981.

Peter C. Fishburn. Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty*, 4 (2):113–134, 1991. doi: 10.1007/BF00056121.

Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. Argumentative large language models for explainable and contestable claim verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14930–14939, 2025. doi: 10.1609/aaai.v39i14.33637.

Jordi Ganzer-Ripoll, Natalia Criado, Maite López-Sánchez, Simon Parsons, and Juan A. Rodriguez-Aguilar. Combining social choice theory and argumentation: Enabling collective decision making. *Group Decision and Negotiation*, 28(1):127–173, 2019. doi: 10.1007/s10726-018-9594-6.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Kaggle. Pokémon – Weedle's Cave: Pokémon battle outcomes. Kaggle dataset, https://www.kaggle.com/datasets/terminus7/pokemon-challenge, 2017. Battle outcomes generated by a simplified battle algorithm, as stated by the dataset author.

Toshihiro Kamishima. Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 583–588, 2003. doi: 10.1145/956750.956823.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. BoardgameQA: A dataset for natural language reasoning with contradictory information. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39052–39074, 2023.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. BIG-Bench Extra Hard. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 26473–26501, 2025. doi: 10.18653/v1/2025.acl-long.1285.

John G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

Nicholas Mattei and Toby Walsh. PrefLib: A library for preferences. In *Algorithmic Decision Theory (ADT)*, volume 8176 of *Lecture Notes in Computer Science*, pp. 259–270, 2013. doi: 10.1007/978-3-642-41575-3_20.

Kenneth O. May. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica*, 22 (1):1–13, 1954. doi: 10.2307/1909827.

Sanjay Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9–10):901–934, 2009. doi: 10.1016/j.artint.2009.02.001.

Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013. doi: 10.1016/j.artint.2012.10.008. Corrected version: arXiv:1804.06763.

Sanjay Modgil and Henry Prakken. The ASPIC$^+$ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014. doi: 10.1080/19462166.2013.869766.

Sanjay Modgil and Henry Prakken. Corrigendum to "A general account of argumentation with preferences" [Artif. Intell. 195 (2013) 361–397]. *Artificial Intelligence*, 263:107–110, 2018. doi: 10.1016/j.artint.2018.05.001.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768, 2024.

Donald Nute. Defeasible logic. In *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3*, pp. 353–395. Oxford University Press, 1994. doi: 10.1093/oso/9780198537472.003.0007.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Nico Potyka. Interpreting neural networks as quantitative argumentation frameworks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6463–6470, 2021. doi: 10.1609/aaai.v35i7.16801.

Henry Prakken and Giovanni Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-Classical Logics*, 7(1–2):25–75, 1997. doi: 10.1080/11663081.1997.10510900.

Qwen Team. Qwen2.5 technical report. arXiv:2412.15115, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. Discontinuity-free decision support with quantitative argumentation debates. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pp. 63–73, 2016.

Patrick Slater. Inconsistencies in a schedule of paired comparisons. *Biometrika*, 48(3/4):303–312, 1961. doi: 10.2307/2332752.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47345–47377, 2024.

Amos Tversky. Intransitivity of preferences. *Psychological Review*, 76(1):31–48, 1969. doi: 10.1037/h0026750.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. HelpSteer2-Preference: Complementing ratings with preferences. arXiv:2410.01257, 2024a.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 1474–1501, 2024b.

Hugh E. Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968. doi: 10.1090/S0002-9947-1968-0226281-1.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. In *International Conference on Learning Representations (ICLR)*, 2025.

Ernst Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929. doi: 10.1007/BF01180541.

A PROOFS

A.1 LEMMA 1 (TRANSITIVITY BARRIER)

$x \succ x$ would need $\neg(x \succcurlyeq x)$, contradicting reflexivity. From $x \succ y$ and $y \succ z$ we get $x \succcurlyeq y \succcurlyeq z$, so $x \succcurlyeq z$; if also $z \succcurlyeq x$ then $z \succcurlyeq x \succcurlyeq y$ gives $z \succcurlyeq y$, contradicting $y \succ z$; so $x \succ z$. A transitive irreflexive relation has no cycle. □

A.2 THEOREM 1 (COMPARATOR SEPARATION)

The $\rho^\star$-preponderance relation on $\{a, b, c\}$ contains the strict 3-cycle $a \succ b \succ c \succ a$ (Figure 1). By Lemma 1 no preorder has a cyclic strict part. □

A.3 THEOREM 2 (EXTENSION SEPARATION)

Under $\rho^\star$ each mutual pair is resolved strictly ($a \succ b$, $b \succ c$, $c \succ a$), so the defeat rule keeps exactly $a \to b$, $b \to c$, $c \to a$. A preorder induces, for each mutual pair, a one-way defeat (strict pair) or a symmetric pair of defeats (indifferent pair); matching the directed 3-cycle requires all three pairs strict and cyclically oriented, which Lemma 1 forbids. Semantics: a directed odd cycle has no non-empty admissible set (an argument must be defended against its unique attacker, which requires that attacker's attacker, and so on around the cycle, so all three would be in, contradicting conflict-freeness), hence grounded $= \emptyset$ and preferred $= \{\emptyset\}$. The conflict-free sets are $\emptyset$ and the three singletons; a stable extension must attack every outside argument, which $\emptyset$ does not, and no singleton does (each attacks only its successor), so no stable extension exists. Any preorder yields either the symmetric clique (all indifferent), with preferred $= \{\{a\}, \{b\}, \{c\}\}$ and three stable extensions, or a defeat graph with at least one strict pair whose strict pairs are consistently oriented by the barrier, which has a non-empty preferred extension. Hence, no preorder matches the (preferred, stable) profile. □

A.4 THEOREM 3 (CYCLES OF EVERY ORDER)

Compare consecutive shifts v_s, v_{s+1} (indices modulo k). Coordinatewise $v_{s+1}[i] = v_s[i] + 1$ except at the single index $i^\star$ with $v_s[i^\star] = k - 1$ and $v_{s+1}[i^\star] = 0$. Thus $|v_s[i] - v_{s+1}[i]| = 1$ for $i \neq i^\star$ and $= k - 1$ at $i^\star$. For $k \geq 3$, $k - 1 > 1$, so $i^\star$ is the unique largest gap; MAX-GAP decides there and $v_s[i^\star] = k - 1 > 0$, so $v_s \succ v_{s+1}$ for every s. □

A.5 THEOREM 4 (CAPACITY OF MAX-GAP)

(a) For $m = 1$, a robust realization is a set of distinct scores, and the induced relation is the linear order of the scores; every linear order arises. For $m = 2$, suppose x beats y with deciding criterion h and other criterion h'. Robustness gives $|d_h| > |d_{h'}|$ and the verdict gives $d_h > 0$, so $d_h + d_{h'} \geq d_h - |d_{h'}| > 0$: the winner has the strictly larger coordinate sum $s_1 + s_2$. The strict relation is therefore contained in the strict order by coordinate sum, which is acyclic, so every robustly realized tournament is transitive; conversely, scores $s(v) = (r(v), 0)$ with distinct r robustly realize any linear order (the gap on the second criterion is 0, so the first is the unique maximum).

(b) Intransitive tournaments are realized for $m = 3$ by the witness. For $m = n$, index criteria by vertices. Given a tournament T, put $s_{h_v}(v) = B = 3$ and, for $u \neq v$, $s_{h_v}(u) = 1$ if $u \to v$ in T and 0 otherwise. For a pair (u, w) with $u \to w$: on h_u the gap is $|B - s_{h_u}(w)| = B - 0 = 3$ because w does not beat u; on h_w it is $|s_{h_w}(u) - B| = B - 1 = 2$; on any other h_x both scores lie in $\{0, 1\}$ so the gap is at most 1. The maximum is unique at h_u, where $s_{h_u}(u) = 3 > 0 = s_{h_u}(w)$, so u wins. Hence, every pair is oriented as in T with a unique maximum gap. (Checked exhaustively for $n \leq 5$ and on random tournaments up to $n = 12$.)

(c) Fix n and m and regard the $N = nm$ scores as real variables. Under a robust realization the tournament is determined by the signs of the polynomials $d_h(u, v) = s_h(u) - s_h(v)$ (degree 1; $m\binom{n}{2}$ of them) and $d_h(u, v)^2 - d_{h'}(u, v)^2$ (degree 2; $\binom{m}{2}\binom{n}{2}$ of them). Robustness makes the polynomials that decide the verdicts nonzero (the largest gap is unique and its d_h is nonzero), but zero differences and tied smaller gaps may occur; since the deciding inequalities are strict, a sufficiently

small generic perturbation of the scores makes every polynomial nonzero while preserving every unique maximum and every verdict. Hence, each robustly realizable tournament is realized by a strict sign pattern of the full family, and it suffices to count strict sign patterns. By Warren's theorem (Warren, 1968, Thm. 3), P real polynomials of degree at most δ in N variables, with $P \geq N$, realize at most $(4e\delta P/N)^N$ strict sign patterns. Here $P = \binom{n}{2}\binom{m+1}{2} = n(n-1)m(m+1)/4$, $\delta = 2$ and $P \geq N$ whenever $(n-1)(m+1) \geq 4$, so the number of robustly realizable tournaments is at most $(8eP/N)^N = \big(2e(n-1)(m+1)\big)^{nm}$. Dividing by $2^{\binom{n}{2}}$ gives, for fixed m, an exponent $nm\log_2(2e(n-1)(m+1)) - n(n-1)/2 \to -\infty$. Universality needs $\big(2e(n-1)(m+1)\big)^{nm} \geq 2^{n(n-1)/2}$, i.e. $m\log_2(2e(n-1)(m+1)) \geq (n-1)/2$; since $m \leq n$ suffices by (b), the logarithm is $O(\log n)$ and $m = \Omega(n/\log n)$. □

Remark 1 (Unrestricted per-pair policies). Call $\mathcal{A}$ mutually dominating if for every pair each argument beats the other on some criterion. Then every tournament is realized by *some* endogenous policy: for each pair oriented $a \succ b$, pick a criterion on which a is higher and put it first. The policy is chosen pair by pair, so this says only that the policy class is unconstrained; Theorem 4 is the statement about a fixed natural rule.

A.6 Theorem 5 (transitivity ceiling) and Corollary 1

Let L be a linear order on V and let A_L be the set of majority arcs $u \to v$ of T (i.e. $n_{uv} > n_{vu}$) with v ranked above u in L. The outcomes L predicts wrongly are, for every unordered pair, the outcomes in the direction opposite to L: $\sum_{i<j}\min(n_{ij}, n_{ji})$ from the minority sides, plus $|n_{uv} - n_{vu}|$ for every arc in A_L. Hence $\mathrm{errors}(L) = L_{\mathrm{pair}} + w(A_L)$. The set A_L is a feedback arc set: after removing it every remaining arc points downward in L, so no directed cycle survives. Conversely, if F is any feedback arc set, $T - F$ is acyclic and has a topological order L_F, and every arc that disagrees with L_F lies in F, so $w(A_{L_F}) \leq w(F)$. Therefore $\min_L w(A_L) = \mathrm{FAS}_w(T)$ and the maximum number of correct predictions is $N - L_{\mathrm{pair}} - \mathrm{FAS}_w(T)$. A model whose verdicts are induced by a weak order with any deterministic tie-breaking is a linear order, so the same bound applies; a pairwise predictor can take the majority side of every pair and is bounded only by $N - L_{\mathrm{pair}}$.

For the corollary, let T be a complete tournament with c_3 cyclic triangles. Every cyclic triangle contains at least one arc of A_L (three arcs all agreeing with a linear order cannot form a cycle), and every arc lies in exactly $n-2$ triangles, so $c_3 \leq (n-2)|A_L|$ and the accuracy is at most $1 - c_3/((n-2)\binom{n}{2}) = 1 - c_3/(3\binom{n}{3}) = 1 - f_3/3$. □

A.7 Theorem 6 (sufficient condition; failure under combining rules)

(a) By Modgil & Prakken (2013, Def. 17), a strict continuation A of $\{C_1, \ldots, C_n\}$ has $\mathrm{Prem}_p(A) = \bigcup_i \mathrm{Prem}_p(C_i)$ and $\mathrm{DefRules}(A) = \bigcup_i \mathrm{DefRules}(C_i)$. If every strict rule has one antecedent, every argument is a chain whose fallible material (ordinary premises and defeasible rules) is that of a single maximal fallible sub-argument; for $n \geq 2$ independent C_i (pairwise disjoint, nonempty fallible material) the required union cannot be the fallible material of a chain, so no strict continuation of the set exists and the universally quantified antecedent of clause 18-2 is unsatisfiable. Clause 18-1 holds by the strength gate with CERTAIN encoded as strict-and-firm and scores inherited by strict continuations; asymmetry holds because ρ depends on the unordered pair. So the ordering is reasonable and Corollary 2 gives the postulates. (b) Add strict rules $\mathrm{Conc}(b), \mathrm{Conc}(c) \Rightarrow x_{bc}$, $\mathrm{Conc}(a), \mathrm{Conc}(c) \Rightarrow x_{ac}$, $\mathrm{Conc}(a), \mathrm{Conc}(b) \Rightarrow x_{ab}$, giving continuations X_{bc}, X_{ac}, X_{ab}. An ordering with $X_{bc} \prec a$, $X_{ac} \prec b$, $X_{ab} \prec c$ violates clause 18-2 for $\{a, b, c\}$. □

A.8 Theorem 7 (aggregation decides the set clause)

Write $A_{ij} = s_{i,h_j}$ for the score of argument i on criterion h_j and $G_i = g(\{s_j\}_{j\neq i})$ for the aggregate the continuation of the others receives.

(a) Suppose, for contradiction, that every C_i strictly beats G_i under MAX-GAP with fixed tie-break π, and let h_i be the deciding criterion of that comparison (the first criterion in the MAX-GAP order, so $\mathrm{gap}_i(h_i) \geq \mathrm{gap}_i(h)$ for all h, with h_i preceding h in π when the gaps tie), where $\mathrm{gap}_i(h) = |s_{i,h} - G_{i,h}|$. Winning on h_i means $s_{i,h_i} > G_{i,h_i} \geq \max_{j\neq i} s_{j,h_i}$ by max-domination, so i is the unique strict maximum on h_i and the map $i \mapsto h_i$ is injective. Take $i \neq j$ (the clause is trivial for

$n = 1$). Since j is the unique maximum on h_j and $G_{i,h_j} \geq s_{j,h_j} > s_{i,h_j}$, the criterion h_j is a losing criterion for i with

$$\mathrm{gap}_i(h_j) = G_{i,h_j} - s_{i,h_j} \;\geq\; A_{jj} - A_{ij} > 0, \quad \text{while} \quad \delta_i := \mathrm{gap}_i(h_i) = s_{i,h_i} - G_{i,h_i} \;\leq\; A_{ii} - A_{ji}.$$

MAX-GAP requires $\delta_i \geq \mathrm{gap}_i(h_j)$, hence $A_{ii} - A_{ji} \geq A_{jj} - A_{ij}$; exchanging i and j gives the reverse inequality. All inequalities are therefore equalities, in particular $\delta_i = \mathrm{gap}_i(h_j)$ and $\delta_j = \mathrm{gap}_j(h_i)$. With tied gaps, the verdict for i requires h_i to precede h_j in π and the verdict for j requires h_j to precede h_i: impossible for one fixed order. So some C_i does not beat G_i, and the clause holds. This holds for any choice of continuation, since every continuation of the others receives the same aggregate by Definition 17.

(b) On the k-cycle witness, for each i let h be the criterion with $C_{i,h} = k - 1$. The other arguments take the values $0, \ldots, k-2$ there, so $\min_{j \neq i} s_{j,h} = 0$ and, by min-domination and nonnegativity, $G_{i,h} = 0$: the gap is $k - 1$. On any other criterion h', $C_{i,h'} \leq k - 2$ and $G_{i,h'} \in [0, 1]$ (the minimum over the others is 0, or 1 when $C_{i,h'} = 0$), so the gap is at most $\max(k - 2, 1) = k - 2$ for $k \geq 3$. The maximum gap is unique at h and C_i wins there; every C_i beats its continuation, and the clause is violated. □

A.9 PROPOSITION 1 (FLIP CONDITIONS)

(a) The pair is PROBABLE–PROBABLE, so the verdict is the sign of d on the first discriminating criterion. After the swap, the set of criteria on which the two arguments differ is unchanged, and the sign on $h^\star$ is reversed. If $h^\star$ is again the first discriminating criterion of the (possibly changed) ordering, the verdict reverses. Otherwise some other discriminating criterion h' comes first and the verdict is the sign of the unchanged $d_{h'}$, which may or may not agree with the original verdict, so the condition is sufficient but not necessary: with $a = (2, 0)$, $b = (0, 1)$ and the symmetric policy "order $(1, 2)$ if $|\sum s(a) - \sum s(b)| \leq 1$, else $(2, 1)$", criterion 1 decides for a; after the swap $a' = (0, 0)$, $b' = (2, 1)$, the policy switches to $(2, 1)$, criterion 2 decides for b', and the verdict reverses although the deciding criterion changed. (b) Under swap-invariance, the ordering is unchanged, the criteria before $h^\star$ still tie, and $h^\star$ still discriminates. (c) With total scores $8 > 1$ the policy puts first the criteria on which a is higher: h_0 ($2 > 1$) decides for a. After swapping on h_0, $a = (1, 1, 5)$, $b = (2, 0, 0)$, totals $7 > 2$, and the first criterion on which a is higher is h_1 ($1 > 0$): a still wins. (d) With strengths and all non-intervened terms fixed, the gate depends on $|d|$ and is unchanged by the swap; $\bar{w}$ is therefore unchanged, and $\kappa_{h^\star}$ changes sign, so $z \mapsto z - 2\kappa_{h^\star}$. The strict signs of z and $z - 2\kappa_{h^\star}$ are opposite iff their product $z(z - 2\kappa_{h^\star})$ is negative, iff $|\kappa_{h^\star}| > |z - \kappa_{h^\star}|$; when the product is zero, one of the logits vanishes, and the model is tied rather than reversed. □

B APPLICATION OF THE ASPIC$^+$ POSTULATE THEOREMS

Corollary 2 applies Theorems 12–15 of Modgil & Prakken (2013) (sub-argument closure, closure under strict rules, direct consistency, indirect consistency). We read the corrected version of the paper (arXiv:1804.06763), which the authors state "contains corrections to errors in the original"; the corrigendum (Modgil & Prakken, 2018) reports that all definition and proposition numbers are unchanged and that all results are unaffected by the changes.

Hypotheses. (H1) The (c-)structured argumentation framework is *well defined* (Def. 12): axiom consistent, well formed (a contrary of a formula is not an axiom nor the consequent of a strict rule), and closed under contraposition *or* transposition (c-classical, for c-SAFs). (H2) The argument ordering is *reasonable* (Def. 18): 18-1(i) a strict-and-firm argument is strictly preferred to every plausible or defeasible argument; 18-1(ii) a strict-and-firm argument is never strictly less preferred than any argument; 18-1(iii) for a strict continuation A' of $\{A\}$, $A \not\prec B \Rightarrow A' \not\prec B$ and $B \not\prec A \Rightarrow B \not\prec A'$; 18-2 the strict part is asymmetric and for any finite $\{C_1, \ldots, C_n\}$ and strict continuations $C^{+\setminus i}$ of $\{C_1, \ldots, C_n\} \setminus \{C_i\}$ it is not the case that $C^{+\setminus i} \prec C_i$ for all i. (H3) Theorems 12, 13 and 15 concern attack-complete extensions and Theorem 14 attack-admissible extensions; arguments are finite. No clause requires transitivity; asymmetry forbids 2-cycles but not 3-cycles.

Mapping. The structured FULCRUM framework instantiates a propositional language with a contrariness relation, an axiom, and ordinary premises, strict and defeasible rules, arguments built by forward chaining, and rebut and undermine attacks; well-definedness is checked per theory. A CERTAIN

argument is an argument built from axiom premises by strict rules only (strict-and-firm); a PROBABLE argument contains an ordinary premise or a defeasible rule (fallible). The ordering: strict-and-firm above fallible (18-1(i)–(ii)); two strict-and-firm arguments equipollent; two fallible arguments compared by preponderance on their scores. *Score inheritance.* The score of a fallible argument is the score vector attached to its top defeasible rule, or to its ordinary premise if it is a premise argument (its maximal fallible sub-argument in the sense of Def. 16 is itself); a strict continuation of a single argument receives the same vector, which gives 18-1(iii) because the comparator sees identical inputs; a strict continuation of a set receives the aggregate g of its maximal fallible sub-arguments' vectors (Section 4.2). The strict ordering $\prec$ is defined directly from the verdicts, as in the corrected source, which defines the strict argument ordering directly rather than as the strict part of a non-strict one. Asymmetry holds because ρ is a function of the unordered pair.

What the corrigendum changed. A counter-example due to Dyrkolbotn, also cited by Dung, showed that an ordering derived from the Elitist set comparison of a preordering on premises could fail clause 18-2 and violate consistency, although the set comparison satisfied the original definition of "reasonable inducing" (Def. 22). The corrected paper defines the Elitist and Democratic strict set comparisons directly from the strict base ordering (Def. 19), defines the strict argument ordering directly (Defs. 20–21), and revises Def. 22 accordingly. These changes concern the *derivation* of orderings from premise and rule orderings, which FULCRUM does not use: its ordering is given directly by the comparator and the question of reasonableness is addressed by Theorems 6–7. The episode confirms that the set clause of 18-2 is the operative condition.

What is and is not claimed. The corollary is an application, not a re-proof. What is new is the construction of an intransitive ordering that satisfies the hypotheses (Theorem 6(a)) and the aggregation conditions that decide the set clause when combining rules are present (Theorem 7); a full reasonableness proof for the combined ordering (weakest link for logical strength together with a max-dominating aggregation for preponderance) in a specific instantiation is not given.

C PROPERTIES OF THE DEFEAT MAP

All properties are proved for the defeat rule of Section 3 and were additionally checked on $4{,}000$ random frameworks under MAX-GAP.

Proposition 2 (Well-definedness)**.** *For a fixed framework and policy, the defeat relation is a deterministic function of $(\mathcal{A}, R, \rho)$, independent of the order in which R is enumerated.*

Proof. Each mutual pair is resolved once; since ρ is a function of the unordered pair, comparing (x, y) returns the mirror of comparing (y, x). One-way attacks are handled pointwise. □

Proposition 3 (Recovery)**.** *If every argument is* PROBABLE *with equal scores and $\rho = \rho_0$, the defeat relation equals R and all extensions coincide with those of $(\mathcal{A}, R)$.*

Proof. Every comparison is equipollent; a mutual pair keeps both directions, and a one-way attack is never repelled. □

Proposition 4 (Mutual conflicts are never dissolved; direct consistency)**.** *Every mutually attacking pair retains at least one defeat direction; hence every grounded, preferred, or stable extension is conflict-free and, if contradiction is encoded as mutual attack, no two contradictory arguments are jointly accepted, even when $\succcurlyeq_\rho$ is intransitive.*

Proof. The resolution returns left, right or equipollent, keeping $\{x \to y\}$, $\{y \to x\}$ or both; a conflict-free set cannot contain both endpoints of a retained edge. Intransitivity appears only as cyclic defeat, which the semantics resolve by non-acceptance. □

Proposition 5 (Complexity)**.** *For $\sigma \in \{$grounded, complete, preferred, stable$\}$, credulous and skeptical acceptance in a* FULCRUM *framework are complete for the same class as in Dung frameworks.*

Proof. Computing the defeat relation costs $O(|R|\, m)$ (one scan per attack); acceptance is then the Dung problem on the defeat graph, giving the upper bounds; by Recovery, every Dung framework

is a FULCRUM instance with trivial preference, so the lower bounds of Dvořák & Dunne (2017) transfer. □

D ADDITIONAL RESULTS AND FRAMEWORK COMPARISON

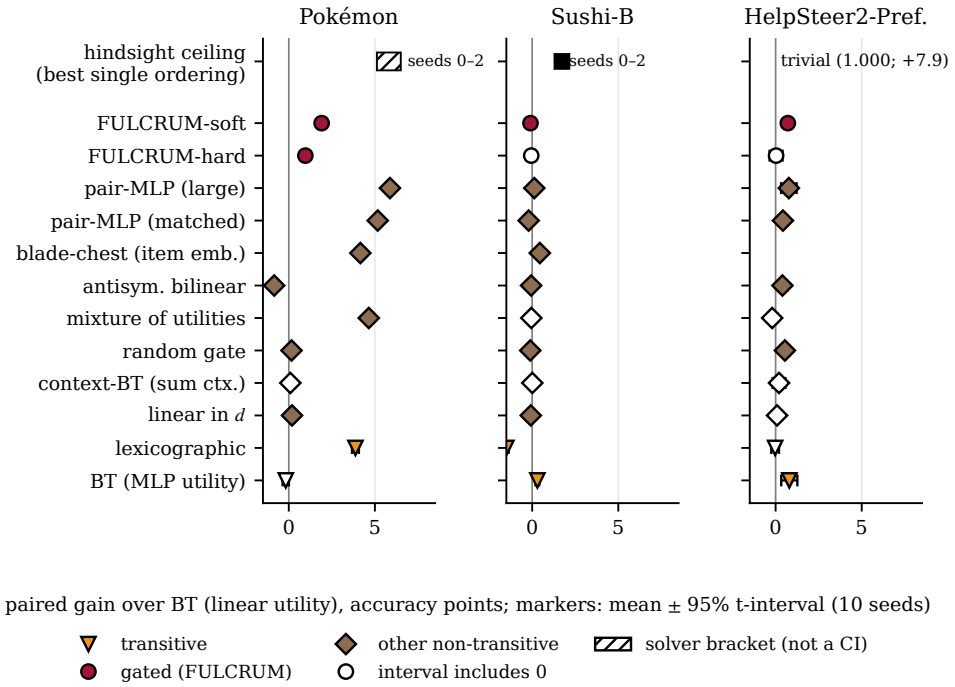


Figure 3: Paired per-seed gain over linear BT in accuracy points with 95% t-intervals over ten seeds (filled markers: interval excludes zero). The top row is the hindsight ceiling of Theorem 5 minus BT, computed pairwise on seeds 0–2: a hatched solver bracket (not a confidence interval) on Pokémon, a value bounded within 0.001 on Sushi-B, and the trivial value 1 on the HelpSteer2 matching. The Condorcet and MAX-GAP rules fall below the axis and are omitted.

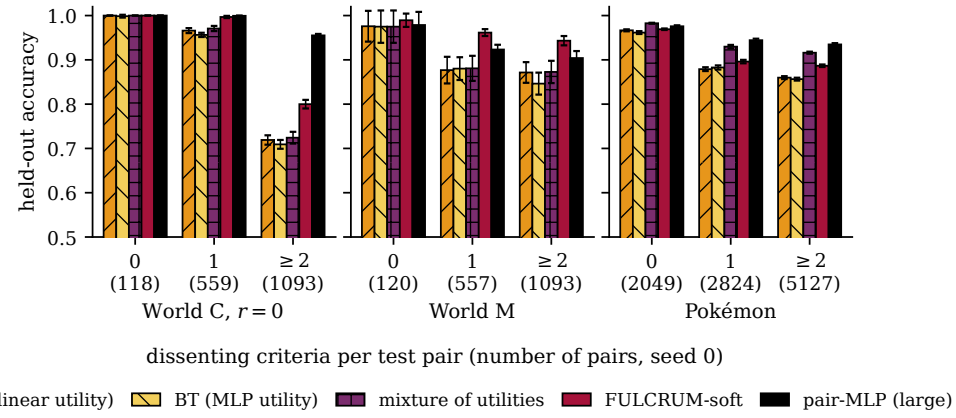


Figure 4: Held-out accuracy by the number of criteria that dissent from the majority on each test pair (pair counts, seed 0, under each bin; ten seeds, t-intervals). The models are nearly indistinguishable where all criteria agree, and the gap between the pair-dependent comparators and the single ordering widens as more criteria dissent, in both controlled worlds and on Pokémon.

Exhaustive checks. All 29 preorders on three elements (count as in OEIS A000798): none realizes the witness comparator table, none induces the directed-cycle defeat graph, none matches its (preferred, stable) profile. With a spoiler $d \to a$ added to the witness, FULCRUM gives grounded $= \{b, d\}$

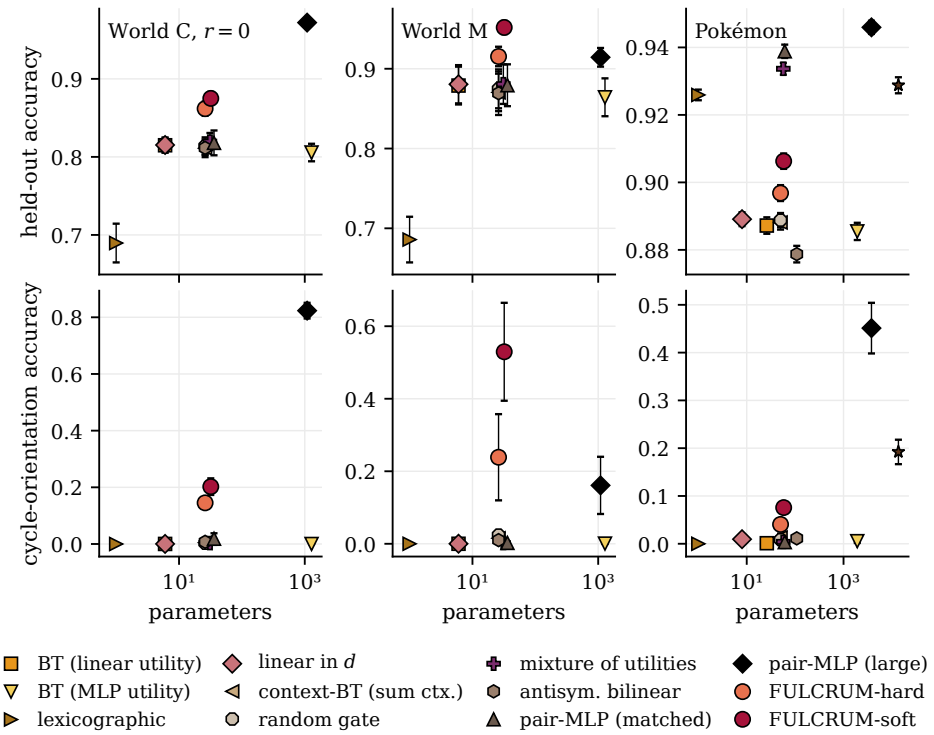


Figure 5: Every learned model by parameter count (log scale) against held-out accuracy (top) and the fraction of cyclic test triples whose orientation it reproduces (bottom). Every transitive model sits at zero orientation (Lemma 1); at the parameter count of a linear utility only the gated comparators orient any cycle, and the large pair-MLP buys its higher accuracy and orientation with thirty to seventy times more parameters.

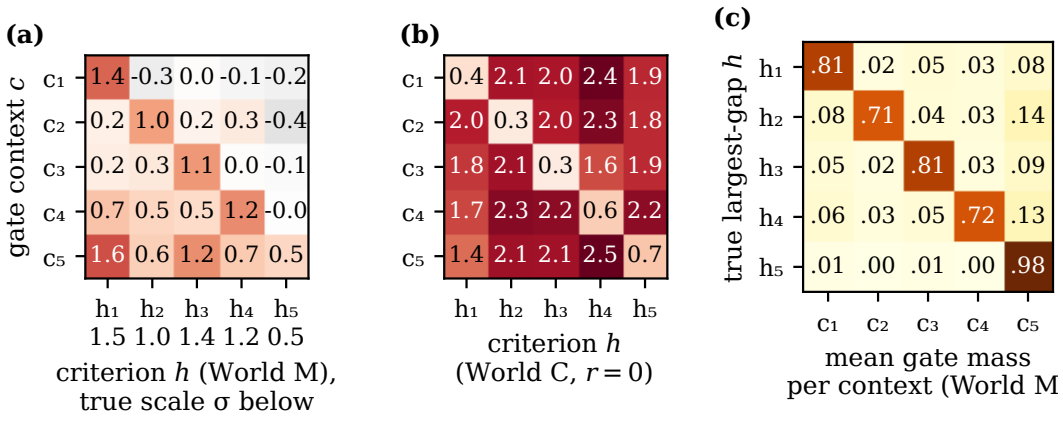


Figure 6: The fitted soft gate (seed 0). (a) Under the MAX-GAP truth of World M, the weight matrix W (rows: gate contexts; columns: criteria) is dominated by its diagonal, whose entries follow the hidden per-criterion scales printed under the columns; (c) pairs whose true largest-gap criterion is h receive most of their gate mass on context c_h. (b) Under the majority oracle of World C, the same architecture learns small diagonal and large off-diagonal weights, discounting the largest-gap criterion and counting the others.

and 79 of the 355 preorders on four elements induce the same grounded extension (grounded does not separate). Theorem 3 is checked for $k = 3, \ldots, 7$. Cyclic triples among distinct score vectors over three criteria: MAX-GAP 44 on $\{0, 1, 2\}$ and 754 on $\{0, \ldots, 3\}$; least-gap-first 364 and largest-sum-first 100 on $\{0, 1, 2\}$. Capacity construction (Theorem 4(b)): all 8, 64, 1024 tournaments for $n = 3, 4, 5$ realized; on the grid $\{0, \ldots, 4\}^m$, MAX-GAP produces exactly $n!$ distinct tie-free tournaments for $m \leq 2$ ($n \leq 5$) and all 8 for $n = 3$, $m = 3$.

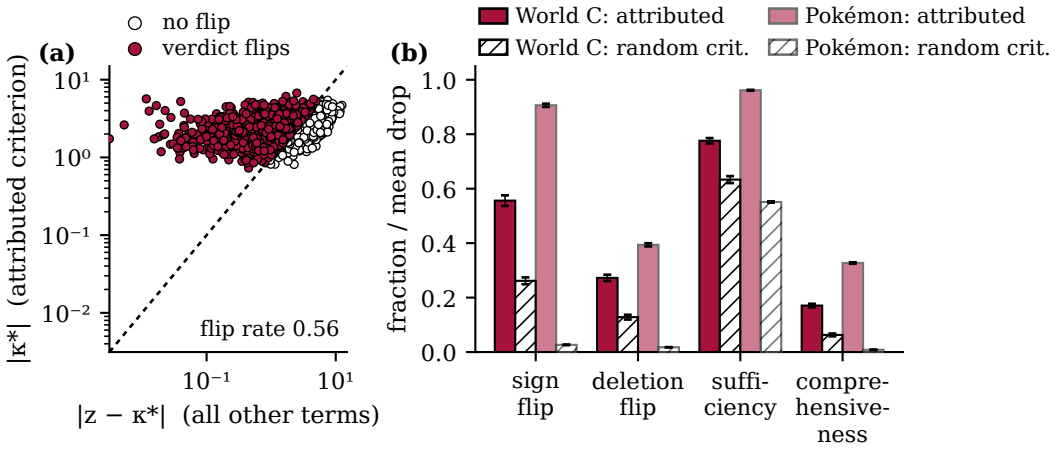


Figure 7: (a) Proposition 1(d) on the held-out pairs of World C ($r = 0$, seed 0): a verdict flips under the score swap exactly when the attributed contribution $|\kappa_{h^\star}|$ exceeds the net of all other terms $|z - \kappa_{h^\star}|$ (points above the diagonal). (b) Four interventions on the attributed criterion (solid) and on a random other criterion (hatched) for World C and Pokémon, ten seeds.

Table 2: Clause 18-2 under conjunction introduction by the rule g that scores a conjunction for preponderance. Middle: the symmetric k-cycle witness (proved for all k, Theorem 7; checked $k \leq 8$). Right: exhaustive count over all 68 preponderance 3-cycles with integer scores in $\{0, \ldots, 3\}$ over three criteria.

aggregation g	k-cycle witness	grid 3-cycles with clause holding
minimum (weakest link)	violated, all k	36/68
mean	holds on the witness	66/68
maximum	holds, all k	68/68
sum	holds, all k	68/68

Aggregation theorem. On 20,000 random finite sets ($n \in \{2, \ldots, 6\}$ arguments, $m \in \{2, \ldots, 6\}$ criteria, integer scores in $\{0, \ldots, 9\}$, random tie-break) the set clause held in every case for $g = \max$, $g = \text{sum}$ and $g = \max + 0.5$, failed in 153 cases for $g = \text{mean}$, and failed on every k-cycle witness for $k = 3, \ldots, 8$ under $g = \min$ and $g = \frac{1}{2}\min$. Proposition 1: the score swap on the deciding criterion reversed all 19,965 non-equipollent random MAX-GAP comparisons tested; the higher-total policy failed on the stated pair.

Table 3: Tag-based discrete study (7 seeds, 1,200 conflicts each, 60/40 split): accuracy ± t-based 95% half-width and paired gap. The preregistered labels are A1 = multi-criterion structure off (one criterion) and A2 = context conditioning off (one ordering).

regime	model	accuracy	gap vs. full (paired)	flip
context-dependent	full (one ordering per tag)	0.898 ± 0.017	–	1.000
	A2: one global ordering	0.656 ± 0.017	-0.242 ± 0.016	–
	A1: one criterion	0.627 ± 0.013	-0.271 ± 0.015	–
transitive (control)	full / A1 / A2	0.900 ± 0.015	0.000	1.000

Tag-based discrete study. With a ground truth in which the deciding criterion is a function of a per-conflict tag with three values (label noise 0.1, 1,200 conflicts, 60/40 split, seeds $\{11, \ldots, 77\}$), fitting one criterion ordering per tag reaches 0.898 ± 0.017; one global ordering 0.656 ± 0.017; one criterion 0.627 ± 0.013 (paired gaps 0.242 ± 0.016 and 0.271 ± 0.015); in a transitive world all three coincide at 0.900 ± 0.015; the discrete flip diagnostic is 1.000. Because the context is an exogenous tag, this study is a check of the fitting code rather than evidence about endogenous pair-dependence, and we report it here only.

Strength gate and calibration. On constructed heterogeneous-noise data (2,400 conflicts per seed, 35% CERTAIN arguments, noise 0.02 on CERTAIN-involved and 0.25 on PROBABLE–PROBABLE conflicts, 5 seeds, both models temperature-scaled): with the gate, accuracy 0.823, Brier 0.125, confidence 0.97 and Brier 0.025 on CERTAIN-involved conflicts, confidence 0.72 on PROBABLE

conflicts; without it, accuracy 0.609, Brier 0.238, confidence flat at 0.60 and Brier 0.260 on CERTAIN-involved conflicts; aggregate ECE ties (0.043 vs 0.039) because temperature scaling equalizes it. The near-certain confidence in CERTAIN-involved conflicts is partly by construction.

Table 4: Intransitivity in the studied data, computed from the raw files. A triple is decisive if every pair has a strict majority; the fraction is over decisive triples.

data	unit	fraction cyclic
Netflix (PrefLib 00004), 150 strict-order profiles	profiles with a cycle	0.000
Sushi-A majority, 10 items (5,000 rankings)	triples	0.000
Sushi-B majority, 100 items (pairs with ≥ 20 voters)	triples	0.0108 (391/36241)
Sushi, majority of four item attributes	triples	0.0009 (76/81194)
World C oracle, $r = 0$ (Table 5)	triples	0.067
Pokémon 18-type chart	type triples	0.328 (42/128)
Pokémon pairs, dual-type advantage (20k samples)	triples	0.215

E BOARDGAMEQA: FIXED-PRIORITY COMPATIBILITY

BoardgameQA (Kazemi et al., 2023) asks whether a statement is proved, disproved, or unknown given facts, defeasible rules, and priorities over rules stated in natural language; every conflict carries one explicit priority, so FULCRUM and ASPIC$^+$ coincide, and the experiment tests compatibility only. We parse the structured `theory` field deterministically and solve by defeasible logic with team defeat, with no language-model formaliser. On the first 400 test items per split, accuracy (macro-F1) is: KnowledgeLight-depth2 0.810 (0.813), Main-depth1 0.715 (0.715), Main-depth2 0.623 (0.616), ZeroConflict-depth2 0.620 (0.613), EasyConflict-depth2 0.615 (0.603), Main-depth3 0.552 (0.530), KnowledgeHeavy-depth2 0.427 (0.347), Binary-depth3 0.302 (0.295; the Binary splits have no "unknown" label, so the solver's "unknown" outputs are scored as wrong). Accuracy tracks world-knowledge load and degrades with depth; 98% of errors are items where matching a rule antecedent needs commonsense. On a balanced 60-item subset of Main-depth1, the solver scores 0.733 (macro-F1 0.729) and Qwen2.5-7B-Instruct (Qwen Team, 2024) with greedy decoding and no chain of thought, run once, scores 0.567; over-prediction of "unknown" has also been reported for much larger models (Kazemi et al., 2025). LLM-ASPIC$^+$ (Fang et al., 2025), which adds a language-model formaliser, reports 87.1% on depth-2 and 82.6% on depth-3 with a fixed superiority relation; it lifts the world-knowledge ceiling our no-LLM solver hits and is orthogonal to pair-dependent priorities. We also asked the same model, with 24 in-context examples, to resolve 40 conflicts from the tag-based generator; it scored 0.425 against 0.875 for the fitted discrete calculus, showing only that this abstract rule was not recovered by in-context learning. These probes are single runs and are not competitive language-model evaluations.

F EXPERIMENTAL DETAILS AND FULL RESULTS

All experiments are deterministic given the seed. Models were trained on one GPU (Quadro RTX 8000) with PyTorch (Paszke et al., 2019); the exhaustive checks and the ceiling solver run on a CPU. Every number in the paper is derived from the per-seed results supplied as supplementary material.

Models. Every model receives, for a pair (a, b), the per-item feature vectors $x_a, x_b \in \mathbb{R}^k$, the criterion-difference vector $d \in \mathbb{R}^m$ (which may contain pair-level criteria), item identifiers where items recur, and the label. Every model has an antisymmetric score $z(a, b) = -z(b, a)$ and is evaluated as $z(a, b) + b_0$ with one shared scalar position intercept b_0. A model is called transitive when its score is a utility difference, $z = u(x_a) - u(x_b)$; its score then induces a single ordering of the items, whereas the thresholded predictor $z + b_0$ does so only when $b_0 = 0$. *Transitive:* BT (linear utility) $z = w \cdot (x_a - x_b)$; BT (MLP utility) $z = f(x_a) - f(x_b)$ with f a two-hidden-layer tanh network of width 32; lexicographic (best of the $m!$ criterion orders on the training set, first differing criterion decides, learned polarity). *Non-transitive:* Condorcet rule (majority of criteria with learned polarities and one scale); MAX-GAP rule (fixed gate and weights: the largest-gap criterion decides with learned polarity and one scale); linear in d ($z = w \cdot d$; transitive when d consists of item-feature differences, as in the controlled worlds and Sushi-B, and non-transitive only when d

contains a pair-level criterion, as the type-advantage term on Pokémon); context-BT with a non-gap context (the criterion with the largest combined level $|x_{a,h} + x_{b,h}|$); random gate (a hash of the pair selects the context); mixture of m linear utilities with constant mixture weights; antisymmetric bilinear $z = w \cdot d + x_a^\top (UV^\top - VU^\top) x_b$ with rank-2 factors (feature-based blade-chest); blade-chest with free item embeddings of dimension 8 plus $w \cdot d$ (Chen & Joachims, 2016), where items repeat; antisymmetrised pair-MLP $z = f([x_a, x_b, d]) - f([x_b, x_a, -d])$ with one tanh hidden layer of width h, where the matched version sets h so that the parameter count is closest to $m^2 + 1$ (the hard gate's) and the large version uses $h = 64$; FULCRUM-hard and FULCRUM-soft as in Section 5 (β initialised at 2, gate bias initialised at 0, both learned; variants with fixed $\beta = 4$ and without gate bias in the ablations). Parameter counts are in Table 13.

Training and calibration. Criterion differences and per-item features are standardised by their mean absolute value on the training split (so that $|d_h|$ are comparable across criteria for the gates). Full-batch Adam, learning rate 0.05, weight decay 10^{-4}, binary cross-entropy on the logit, up to 400 epochs (controlled worlds) or 600 epochs (empirical data) with early stopping on validation NLL (patience 50 epochs), then post-hoc temperature scaling on the validation split (Guo et al., 2017). The mixture model is trained on its mixture probability; the lexicographic model is fitted by training accuracy. All models of a seed see identical splits and are evaluated on identical test pairs. A longer schedule (1,500 epochs, no early stopping) changes Pokémon accuracies by at most one point and no ranking.

Metrics and intervals. Accuracy, NLL, Brier score and ECE (15 equal-width confidence bins). Structure metrics use a reference relation over test items (the oracle in the controlled worlds; the majority of observed test outcomes per pair on Pokémon and Sushi-B): accuracy on pairs that lie on at least one cyclic triple versus the rest; cycle-orientation accuracy (fraction of cyclic triples of the reference relation whose three predicted orientations all match it, where a pair's predicted orientation is the majority of the evaluated predictor's probabilities over all test presentations of that pair, in either order); accuracy by criterion-conflict level (number of criteria disagreeing with the criterion majority after polarity alignment). Ten seeds throughout; per-model intervals are t-based ($t_{0.975,9} = 2.262$); gaps are paired per seed, with a paired t-interval and a percentile bootstrap over seeds (10^4 resamples) reported in the result files; "bold" marks paired intervals excluding zero.

Orientation metric and the position intercept. All structure metrics are computed from the evaluated predictor $z(a, b) + b_0$. For a transitive model, z is a utility difference and the relation induced by z alone is acyclic, so its cycle-orientation accuracy is 0 by Lemma 1. The intercept b_0 shifts the decision threshold, and on Pokémon, where a pair can be presented in both orders, the majority orientation of $z + b_0$ over both presentations need not coincide with the orientation of z; this is why Table 11 reports cycle-orientation values of 0.001 (BT) and 0.005 (BT-MLP) on Pokémon rather than exactly 0. The reported values are therefore the orientation of the predictor as evaluated, not of the underlying ordering, for every model; recomputing them from z alone would set the transitive rows to 0 and could change the other rows slightly, and was not done for this version.

World C (Condorcet oracle). $n = 60$ items with $m = 5$ criteria, $x_c = \sqrt{r}\, z + \sqrt{1 - r}\, \varepsilon_c$ with a shared latent z; i beats j iff i is higher on a majority of criteria (ties dropped). Train, validation and test use disjoint item sets (all decisive pairs of each set, about 1,770 per set). Sweeps: $r \in \{1, 0.7, 0.4, 0.2, 0\}$; label noise $\varepsilon \in \{0.1, 0.2, 0.3\}$ applied to train and validation labels only (test labels are clean, so accuracies are not bounded by a noise ceiling) at $r = 0$; training fraction $\{0.1, 0.25, 0.5\}$ of the training pairs at $r = 0$; criterion-identity shuffle (a random permutation of the criterion axes drawn per pair and applied to both items and to d, in train, validation and test).

World M (MAX-GAP truth). $n = 60$, $m = 5$ independent standard normal criteria; hidden scales σ_h log-uniform on $[1/2, 2]$; i beats j by MAX-GAP on $\sigma \odot x$; label noise 0.1 on train and validation labels only (test labels clean); training fractions $\{0.1, 0.25, 1\}$.

Ceilings. For every controlled test tournament we solve the linear-ordering integer program exactly (variables x_{ij}, triangle constraints $x_{ij} + x_{jk} - x_{ik} \in [0, 1]$ added lazily until none is violated; HiGHS via SciPy; at most a minute per instance) and report $1 - \mathrm{FAS}/\binom{n}{2}$, the cyclic-triple fraction f_3 and $1 - f_3/3$. Ceilings for the empirical datasets are computed on the test graphs of seeds 0–2 only

(models are evaluated on all ten seeds; on seeds 0–2 the ceiling and the models share the same test split). For Sushi-B the cycle-covering formulation with lazily generated cycle constraints is solved to optimality on two of the three seeds and bounded within 0.001 on the third; the displayed value 0.672 is the mean of the three upper bounds (the lower bounds agree to three decimals). For HelpSteer2 each response appears in exactly one comparison, so any hindsight ordering satisfies every observed comparison and the ceiling is trivially 1; it says nothing about intransitivity in the underlying population. For Pokémon the test graph has about 10,000 comparisons over 784 items; we report a bracket: the accuracy of an explicit ordering found by insertion local search from the win–loss score (achievable, hence a lower bound on the ceiling) and one minus the value of the LP relaxation of the cycle-covering formulation with lazily generated cycle constraints (a relaxation, hence an upper bound), each after a fixed time budget on seeds 0–2; the displayed bracket is the mean lower and mean upper bound over those seeds. The "ceiling minus BT" row of Figure 3 is computed pairwise on seeds 0–2 only (ceiling bound minus the BT accuracy on the same test split, then averaged), not by subtracting differently averaged summaries. All solvers agree with brute force on random instances with up to seven items.

Pokémon. Kaggle "Weedle's Cave": 50,000 observed battle outcomes among 784 Pokémon with six base stats and one or two types; the standard 18-type effectiveness chart. Criteria: six standardised stat differences and the pair-level type-advantage term $\log \mathrm{off}(i, j) - \log \mathrm{off}(j, i)$ (best effectiveness of the attacker's types against the defender's); per-item features for transitive models: six stats and 18 type one-hots. Splits 70/10/20 by battle, ten seeds. Slices: cross-type battles (non-zero type term) and the lowest 30/20/10% of |total base-stat difference|. Distinct pairs 46,226; 3,573 pairs repeat and 110 of them have inconsistent outcomes; the first-listed Pokémon wins 47.2% of battles (absorbed by b_0).

Sushi-B. Kamishima (2003): 5,000 respondents each ranked 10 items drawn from 100; 45 pairwise comparisons per respondent (225,000 in total), presentation order randomised. Criteria: style, major group, heaviness/oiliness, eating frequency, normalised price, sale frequency; extra per-item features: 12-way minor-group one-hot. Split 60/20/20 by respondent, ten seeds; the test comparison graph is dense (every pair of the 100 items compared by about 45 respondents in the full data). Population majority over the 100 items (pairs with at least 20 voters): 1.1% of decisive triples cycle.

HelpSteer2-Preference. Wang et al. (2024a): 9,125 preference annotations between two responses to the same prompt with strength in $\{-3, \ldots, 3\}$; per-response ratings of helpfulness, correctness, coherence, complexity and verbosity (0–4) from HelpSteer2 (Wang et al., 2024b), joined by exact (prompt, response) text (0 rows unmatched). Ties (2,007 rows) are excluded. Criteria: the five aspect scores. The released train rows are split 90/10 into train and validation by prompt (seeded); the released validation rows are the fixed test set; seeds vary the train/validation split and initialisation. Each response occurs once, so no cycle statistics exist.

Intransitivity counts. Netflix: the 150 strict-order profiles of PrefLib dataset 00004; a profile is cyclic if its pairwise majority relation has a cycle. Sushi: set A (10 items, 5,000 complete rankings) and set B (100 items) majorities; the attribute-majority rule uses the four numerical attributes with "higher is better". Pokémon: type triples are decisive if every pair has a strict winner; Pokémon-pair triples are sampled (20,000 draws).

Additional validation studies. The tag-based discrete study uses 1,200 conflicts per seed, four criteria with integer scores in $\{0, \ldots, 5\}$, a deciding criterion that is a function of a three-valued tag, label noise 0.1, a 60/40 split and seeds $\{11, 22, \ldots, 77\}$; the calibration study uses the model of Section 5 with the hard gate, 2,400 conflicts per seed, and five seeds. BoardgameQA: first 400 test items per split; the language-model probes use Qwen2.5-7B-Instruct with greedy decoding from a local cache, run once.

F.1 Full result tables

Table 5: World C (Condorcet oracle, $m = 5$, 60 items per split, 10 seeds): held-out pairwise accuracy (mean $\pm$ t-based 95% half-width) as the criteria correlation r decreases (conflict increases). T = transitive model. The Condorcet rule is the oracle itself. Ceilings are exact Kemeny/Slater optima on each test tournament (Theorem 5); f_3 is the cyclic-triple fraction and $1 - f_3/3$ the closed-form bound.

model	$r = 1.0$	$r = 0.7$	$r = 0.4$	$r = 0.2$	$r = 0.0$
BT (linear utility)$^{\mathrm{T}}$	0.998 $\pm$0.001	0.934 $\pm$0.007	0.895 $\pm$0.008	0.857 $\pm$0.008	0.815 $\pm$0.010
BT (MLP utility)$^{\mathrm{T}}$	0.998 $\pm$0.001	0.932 $\pm$0.007	0.886 $\pm$0.010	0.851 $\pm$0.007	0.806 $\pm$0.011
lexicographic$^{\mathrm{T}}$	1.000 $\pm$0.000	0.831 $\pm$0.018	0.772 $\pm$0.027	0.706 $\pm$0.027	0.690 $\pm$0.025
Condorcet rule	1.000 $\pm$0.000	1.000 $\pm$0.000	1.000 $\pm$0.000	1.000 $\pm$0.000	1.000 $\pm$0.000
MAX-GAP rule (fixed)	0.998 $\pm$0.002	0.889 $\pm$0.013	0.817 $\pm$0.015	0.756 $\pm$0.016	0.683 $\pm$0.014
linear in d	0.998 $\pm$0.002	0.934 $\pm$0.008	0.894 $\pm$0.009	0.857 $\pm$0.008	0.815 $\pm$0.010
context-BT (sum ctx.)	0.998 $\pm$0.002	0.933 $\pm$0.008	0.892 $\pm$0.010	0.856 $\pm$0.010	0.811 $\pm$0.011
random gate	0.998 $\pm$0.002	0.933 $\pm$0.007	0.893 $\pm$0.008	0.856 $\pm$0.009	0.814 $\pm$0.011
mixture of utilities	0.999 $\pm$0.001	0.934 $\pm$0.008	0.895 $\pm$0.009	0.859 $\pm$0.010	0.820 $\pm$0.011
antisym. bilinear	0.998 $\pm$0.002	0.931 $\pm$0.007	0.888 $\pm$0.009	0.852 $\pm$0.007	0.812 $\pm$0.009
pair-MLP (matched)	0.999 $\pm$0.001	0.932 $\pm$0.009	0.891 $\pm$0.010	0.852 $\pm$0.012	0.818 $\pm$0.016
pair-MLP (large)	1.000 $\pm$0.000	0.969 $\pm$0.007	0.969 $\pm$0.005	0.970 $\pm$0.006	0.972 $\pm$0.002
FULCRUM-hard	0.998 $\pm$0.002	0.946 $\pm$0.006	0.912 $\pm$0.009	0.885 $\pm$0.009	0.862 $\pm$0.009
FULCRUM-soft	0.998 $\pm$0.002	0.946 $\pm$0.005	0.923 $\pm$0.007	0.898 $\pm$0.007	0.875 $\pm$0.009
FULCRUM-soft, fixed β	0.998 $\pm$0.002	0.945 $\pm$0.004	0.921 $\pm$0.009	0.897 $\pm$0.008	0.872 $\pm$0.008
FULCRUM-soft, no bias	0.998 $\pm$0.002	0.951 $\pm$0.005	0.924 $\pm$0.008	0.900 $\pm$0.008	0.877 $\pm$0.008
transitive ceiling (exact)	1.000 $\pm$0.000	0.964 $\pm$0.005	0.936 $\pm$0.008	0.910 $\pm$0.010	0.883 $\pm$0.009
f_3 (cyclic triples)	0.000 $\pm$0.000	0.010 $\pm$0.003	0.026 $\pm$0.005	0.044 $\pm$0.007	0.067 $\pm$0.007
$1 - f_3/3$	1.000 $\pm$0.000	0.997 $\pm$0.001	0.991 $\pm$0.002	0.985 $\pm$0.002	0.978 $\pm$0.002

Table 6: World M (MAX-GAP ground truth with hidden per-criterion scales, label noise 0.1, 10 seeds): accuracy as the training set shrinks. Sample efficiency in the well-specified regime.

model	10%	25%	100%
BT (linear utility)$^{\mathrm{T}}$	0.874 $\pm$0.025	0.874 $\pm$0.026	0.879 $\pm$0.024
BT (MLP utility)$^{\mathrm{T}}$	0.841 $\pm$0.022	0.848 $\pm$0.019	0.864 $\pm$0.024
mixture of utilities	0.871 $\pm$0.025	0.876 $\pm$0.026	0.881 $\pm$0.025
antisym. bilinear	0.831 $\pm$0.033	0.852 $\pm$0.023	0.869 $\pm$0.027
pair-MLP (matched)	0.870 $\pm$0.026	0.874 $\pm$0.026	0.879 $\pm$0.026
pair-MLP (large)	0.867 $\pm$0.018	0.885 $\pm$0.023	0.914 $\pm$0.012
FULCRUM-hard	0.864 $\pm$0.022	0.890 $\pm$0.024	0.915 $\pm$0.012
FULCRUM-soft	0.884 $\pm$0.018	0.922 $\pm$0.022	0.952 $\pm$0.008
transitive ceiling	0.931 $\pm$0.016	0.931 $\pm$0.016	0.931 $\pm$0.016

Table 7: Ablations (10 seeds): accuracy and paired gap to BT (linear utility) of the same run (t-based 95% half-width; bold when the interval excludes zero). Unless stated, World C at $r = 0$; rows vary one ingredient at a time; shuffle, noise and training-size rows are separate runs with their own BT reference.

variant	accuracy	gap vs BT (paired)
FULCRUM-soft (full)	0.875 ±0.009	**+0.060 ±0.006**
hard gate	0.862 ±0.009	**+0.047 ±0.004**
soft gate, fixed β	0.872 ±0.008	**+0.057 ±0.005**
soft gate, no gate bias	0.877 ±0.008	**+0.062 ±0.005**
random selection (gap gating off)	0.814 ±0.011	-0.001 ±0.003
sum-based selection instead of gap	0.811 ±0.011	-0.004 ±0.005
gate and weights fixed (MAX-GAP rule)	0.683 ±0.014	-0.132 ±0.009
single criterion (lexicographic)	0.690 ±0.025	-0.125 ±0.020
shared weights (linear in d)	0.815 ±0.010	+0.000 ±0.002
criterion identity shuffled: FULCRUM-soft	0.876 ±0.008	**+0.061 ±0.009**
criterion identity shuffled: pair-MLP (large)	0.971 ±0.003	**+0.156 ±0.010**
criterion identity shuffled: BT (linear)	0.816 ±0.010	+0.000 ±0.000
World M, unshuffled: FULCRUM-soft	0.952 ±0.008	**+0.073 ±0.027**
World M, unshuffled: pair-MLP (large)	0.914 ±0.012	**+0.035 ±0.020**
World M, criterion identity shuffled: FULCRUM-soft	0.822 ±0.038	**+0.029 ±0.023**
World M, criterion identity shuffled: pair-MLP (large)	0.803 ±0.034	+0.011 ±0.020
label noise $\varepsilon = 0.1$: FULCRUM-soft	0.868 ±0.009	**+0.053 ±0.010**
label noise $\varepsilon = 0.2$: FULCRUM-soft	0.856 ±0.018	**+0.042 ±0.013**
label noise $\varepsilon = 0.3$: FULCRUM-soft	0.830 ±0.023	**+0.019 ±0.015**
training pairs 10%: FULCRUM-soft	0.838 ±0.014	**+0.028 ±0.012**
training pairs 10%: pair-MLP (large)	0.852 ±0.022	**+0.042 ±0.017**
training pairs 25%: FULCRUM-soft	0.857 ±0.009	**+0.045 ±0.006**
training pairs 25%: pair-MLP (large)	0.914 ±0.018	**+0.101 ±0.018**
training pairs 50%: FULCRUM-soft	0.870 ±0.009	**+0.057 ±0.009**
training pairs 50%: pair-MLP (large)	0.953 ±0.008	**+0.140 ±0.013**

Table 8: Empirical datasets (10 seeds): held-out accuracy and NLL (mean ± t-based 95% half-width; all models temperature-scaled on validation data). T = transitive. Blade-chest requires repeated items and is absent on HelpSteer2.

	Pokémon		Sushi-B		HelpSteer2-Pref.	
model	acc	NLL	acc	NLL	acc	NLL
BT (linear utility)$^{\mathrm{T}}$	0.887 ±0.002	0.365 ±0.005	0.656 ±0.003	0.622 ±0.003	0.921 ±0.002	0.169 ±0.001
BT (MLP utility)$^{\mathrm{T}}$	0.885 ±0.003	0.347 ±0.006	0.659 ±0.003	0.619 ±0.003	0.929 ±0.004	0.159 ±0.003
lexicographic$^{\mathrm{T}}$	0.926 ±0.002	0.264 ±0.004	0.641 ±0.003	0.653 ±0.002	0.920 ±0.000	0.225 ±0.002
Condorcet rule	0.712 ±0.003	0.544 ±0.002	0.622 ±0.003	0.652 ±0.002	0.904 ±0.001	0.227 ±0.001
MAX-GAP rule (fixed)	0.676 ±0.003	0.659 ±0.002	0.616 ±0.003	0.673 ±0.001	0.815 ±0.000	0.585 ±0.001
linear in d	0.889 ±0.002	0.364 ±0.006	0.656 ±0.003	0.623 ±0.003	0.922 ±0.002	0.168 ±0.002
context-BT (sum ctx.)	0.888 ±0.002	0.361 ±0.006	0.656 ±0.003	0.622 ±0.003	0.923 ±0.003	0.162 ±0.003
random gate	0.889 ±0.002	0.365 ±0.006	0.655 ±0.004	0.623 ±0.003	0.926 ±0.002	0.169 ±0.006
mixture of utilities	0.934 ±0.002	0.231 ±0.004	0.656 ±0.003	0.622 ±0.003	0.919 ±0.001	0.161 ±0.003
antisym. bilinear	0.879 ±0.002	0.350 ±0.006	0.656 ±0.003	0.621 ±0.003	0.925 ±0.004	0.165 ±0.004
blade-chest (item emb.)	0.929 ±0.002	0.213 ±0.007	0.661 ±0.003	0.617 ±0.003	–	–
pair-MLP (matched)	0.939 ±0.002	0.243 ±0.005	0.654 ±0.004	0.623 ±0.003	0.925 ±0.002	0.166 ±0.002
pair-MLP (large)	0.946 ±0.002	0.136 ±0.004	0.658 ±0.004	0.620 ±0.003	0.928 ±0.004	0.161 ±0.003
FULCRUM-hard	0.897 ±0.002	0.335 ±0.005	0.656 ±0.003	0.623 ±0.003	0.921 ±0.002	0.177 ±0.002
FULCRUM-soft	0.906 ±0.002	0.311 ±0.004	0.655 ±0.003	0.622 ±0.003	0.928 ±0.002	0.166 ±0.002

Table 9: Empirical datasets (10 seeds): Brier score and ECE (15 bins), mean ± t-based 95% half-width.

	Pokémon		Sushi-B		HelpSteer2-Pref.	
model	Brier	ECE	Brier	ECE	Brier	ECE
BT (linear utility)$^{\mathrm{T}}$	0.100 ±0.001	0.084 ±0.003	0.216 ±0.001	0.009 ±0.002	0.050 ±0.000	0.034 ±0.004
BT (MLP utility)$^{\mathrm{T}}$	0.098 ±0.002	0.066 ±0.003	0.215 ±0.001	0.009 ±0.002	0.047 ±0.001	0.030 ±0.004
lexicographic$^{\mathrm{T}}$	0.069 ±0.001	0.004 ±0.002	0.230 ±0.001	0.006 ±0.004	0.058 ±0.000	0.014 ±0.007
Condorcet rule	0.185 ±0.001	0.053 ±0.002	0.230 ±0.001	0.016 ±0.002	0.066 ±0.000	0.012 ±0.004
MAX-GAP rule (fixed)	0.229 ±0.001	0.103 ±0.004	0.240 ±0.001	0.064 ±0.003	0.190 ±0.001	0.200 ±0.005
linear in d	0.099 ±0.001	0.085 ±0.003	0.217 ±0.001	0.009 ±0.002	0.050 ±0.001	0.033 ±0.002
context-BT (sum ctx.)	0.099 ±0.001	0.081 ±0.003	0.216 ±0.001	0.009 ±0.002	0.050 ±0.001	0.030 ±0.005
random gate	0.100 ±0.001	0.085 ±0.003	0.217 ±0.001	0.009 ±0.002	0.050 ±0.001	0.029 ±0.003
mixture of utilities	0.063 ±0.001	0.048 ±0.003	0.216 ±0.001	0.009 ±0.002	0.049 ±0.001	0.032 ±0.002
antisym. bilinear	0.099 ±0.001	0.059 ±0.003	0.216 ±0.001	0.009 ±0.002	0.049 ±0.001	0.036 ±0.004
blade-chest (item emb.)	0.057 ±0.002	0.023 ±0.002	0.214 ±0.001	0.008 ±0.002	–	–
pair-MLP (matched)	0.062 ±0.001	0.025 ±0.002	0.217 ±0.001	0.010 ±0.002	0.049 ±0.001	0.027 ±0.004
pair-MLP (large)	0.040 ±0.001	0.007 ±0.001	0.215 ±0.001	0.009 ±0.001	0.049 ±0.001	0.032 ±0.003
FULCRUM-hard	0.092 ±0.001	0.075 ±0.003	0.216 ±0.001	0.009 ±0.002	0.052 ±0.000	0.033 ±0.004
FULCRUM-soft	0.083 ±0.001	0.068 ±0.003	0.216 ±0.001	0.009 ±0.002	0.049 ±0.000	0.035 ±0.003

Table 10: Pokémon: accuracy on prespecified test slices (10 seeds; mean ± t-based 95% half-width). Slices: all test battles; cross-type battles (non-zero type-advantage term); battles in the lowest 30/20/10% of |total base-stat difference| (a model-free closeness measure).

model	all	cross-type	closest 30%	closest 20%	closest 10%
BT (linear utility)$^{\mathrm{T}}$	0.887 ±0.002	0.872 ±0.004	0.878 ±0.005	0.880 ±0.005	0.884 ±0.007
BT (MLP utility)$^{\mathrm{T}}$	0.885 ±0.003	0.870 ±0.004	0.865 ±0.004	0.864 ±0.006	0.865 ±0.007
lexicographic$^{\mathrm{T}}$	0.926 ±0.002	0.912 ±0.002	0.915 ±0.003	0.916 ±0.003	0.918 ±0.004
Condorcet rule	0.712 ±0.003	0.710 ±0.004	0.579 ±0.005	0.572 ±0.006	0.564 ±0.009
MAX-GAP rule (fixed)	0.676 ±0.003	0.665 ±0.004	0.509 ±0.005	0.500 ±0.008	0.492 ±0.009
linear in d	0.889 ±0.002	0.874 ±0.003	0.881 ±0.004	0.881 ±0.004	0.888 ±0.005
context-BT (sum ctx.)	0.888 ±0.002	0.874 ±0.004	0.882 ±0.004	0.885 ±0.005	0.890 ±0.006
random gate	0.889 ±0.002	0.874 ±0.004	0.881 ±0.004	0.883 ±0.004	0.889 ±0.004
mixture of utilities	0.934 ±0.002	0.920 ±0.002	0.924 ±0.003	0.925 ±0.003	0.926 ±0.004
antisym. bilinear	0.879 ±0.002	0.864 ±0.004	0.859 ±0.005	0.860 ±0.004	0.864 ±0.007
blade-chest (item emb.)	0.929 ±0.002	0.916 ±0.003	0.915 ±0.004	0.915 ±0.004	0.918 ±0.005
pair-MLP (matched)	0.939 ±0.002	0.925 ±0.002	0.931 ±0.004	0.932 ±0.003	0.932 ±0.004
pair-MLP (large)	0.946 ±0.002	0.937 ±0.003	0.936 ±0.004	0.936 ±0.005	0.937 ±0.005
FULCRUM-hard	0.897 ±0.002	0.879 ±0.003	0.884 ±0.005	0.882 ±0.005	0.885 ±0.008
FULCRUM-soft	0.906 ±0.002	0.890 ±0.003	0.894 ±0.005	0.895 ±0.005	0.897 ±0.005

Table 11: Structure metrics (10 seeds, means): accuracy on test pairs that lie on a cyclic triple of the reference relation (on) vs. the rest (off), and cycle-orientation accuracy (cyc.; fraction of cyclic test triples on which all three predicted orientations are correct). Orientations are majorities of the evaluated predictor's probabilities over both presentation orders of a pair and include the shared position intercept; the ordering induced by a transitive model's score orients no cycle, and the small nonzero values for transitive models on Pokémon arise from the intercept (Appendix F). Reference relation: the oracle (World C) or the observed majority among test comparisons (Pokémon, Sushi-B).

	World C, r=0			Pokémon			Sushi-B		
model	on	off	cyc.	on	off	cyc.	on	off	cyc.
BT (linear utility)$^{\mathrm{T}}$	0.741	0.998	0.000	0.606	0.900	0.001	0.638	0.716	0.000
BT (MLP utility)$^{\mathrm{T}}$	0.731	0.994	0.000	0.579	0.899	0.005	0.642	0.717	0.000
mixture of utilities	0.748	0.998	0.003	0.679	0.945	0.004	0.638	0.716	0.000
antisym. bilinear	0.737	0.997	0.008	0.609	0.891	0.011	0.637	0.717	0.001
blade-chest (item emb.)	–	–	–	0.721	0.938	0.192	0.644	0.716	0.003
pair-MLP (matched)	0.747	0.996	0.019	0.685	0.950	0.003	0.636	0.716	0.001
pair-MLP (large)	0.963	0.994	0.823	0.814	0.952	0.451	0.640	0.716	0.004
FULCRUM-hard	0.808	0.997	0.145	0.638	0.909	0.040	0.637	0.716	0.001
FULCRUM-soft	0.827	0.995	0.202	0.662	0.917	0.076	0.637	0.716	0.000
cyclic test triples / pairs on cycles	2500 / 1309 (seed 0)			134 / 372 (seed 0)			5763 / 3171 (seed 0)		

Table 12: Attribution faithfulness (World C, $r = 0$, and Pokémon; 10 seeds, mean over seeds). Interventions on the attributed criterion $h^\star = \arg\max_h |\kappa_h|$ versus a random other criterion (in parentheses): sign reversal and deletion report the fraction of verdicts that flip; sufficiency the agreement when only $h^\star$ is kept; comprehensiveness the mean drop in confidence after deletion. For the pair-MLP the attributed criterion is the one whose deletion changes the logit most (occlusion), because it has no built-in contributions.

data	model	sign flip	deletion flip	sufficiency	comprehensiveness
World C, r=0	FULCRUM-hard	0.61 (0.25)	0.30 (0.12)	0.80 (0.62)	0.18 (0.05)
World C, r=0	FULCRUM-soft	0.56 (0.26)	0.27 (0.13)	0.78 (0.63)	0.17 (0.06)
World C, r=0	linear in d	0.70 (0.19)	0.35 (0.10)	0.85 (0.59)	0.20 (0.02)
World C, r=0	pair-MLP (large)	0.11 (0.05)	0.03 (0.01)	0.98 (0.97)	0.02 (0.00)
Pokémon	FULCRUM-hard	0.92 (0.03)	0.37 (0.02)	0.96 (0.55)	0.31 (0.00)
Pokémon	FULCRUM-soft	0.91 (0.03)	0.39 (0.02)	0.96 (0.55)	0.33 (0.01)
Pokémon	linear in d	0.92 (0.02)	0.36 (0.01)	0.97 (0.55)	0.29 (0.00)
Pokémon	pair-MLP (large)	0.11 (0.02)	0.04 (0.01)	0.97 (0.95)	0.01 (0.00)

Table 13: Parameter counts and mean fit time (seconds, one GPU, full-batch Adam with early stopping) on World C ($m = 5$) and Pokémon ($m = 7$ criteria, 25 per-item features, 784 items).

model	params (C)	time (C)	params (Pk)	time (Pk)
BT (linear utility)	6	0.1	26	0.4
BT (MLP utility)	1282	0.4	1922	0.7
lexicographic	0	0.1	0	69.5
Condorcet rule	2	0.4	2	0.1
MAX-GAP rule (fixed)	2	0.1	2	0.1
linear in d	6	0.1	8	0.2
context-BT (sum ctx.)	26	0.1	50	0.5
random gate	26	0.1	50	0.4
mixture of utilities	31	0.3	57	1.3
antisym. bilinear	26	0.2	108	1.0
blade-chest (item emb.)	–	–	13608	0.6
pair-MLP (matched)	36	0.4	61	1.6
pair-MLP (large)	1090	0.5	3778	0.5
FULCRUM-hard	26	0.2	50	0.7
FULCRUM-soft	32	0.4	58	1.0

G Extended Related Work

Table 14: Preference mechanisms in argumentation and preference learning (citations in the text). Pair-dependent: the priority over criteria can change with the pair; derived cycle: an intransitive defeat cycle is produced from argument features rather than written by hand; postulates: closure and consistency proved (†: under the conditions of Section 4.2).

framework	multi-criteria	pair-dependent	derived cycle	postulates	learnable
preference-based AF	–	–	–	–	–
value-based AF	per audience	–	–	–	–
ASPIC$^+$	rules, premises	–	–	✓	–
ABA$^+$	–	–	–	✓	–
extended AF	–	–	by hand	–	–
gradual / neural AF	–	–	–	–	✓
Bradley–Terry reward model	–	–	–	n/a	✓
blade-chest, bilinear, pair-MLP	–	implicit	✓	n/a	✓
FULCRUM (this paper)	✓	✓	✓	✓†	✓

Preference in argumentation. Preference-based (Amgoud & Cayrol, 1998), value-based (Bench-Capon, 2003), ASPIC$^+$ (Modgil & Prakken, 2014), ABA$^+$ (Čyras & Toni, 2016), extended AFs (Modgil, 2009) and ranking-based semantics (Amgoud & Ben-Naim, 2013) resolve conflict through one ordering with an acyclic strict part; FULCRUM differs in representation while keeping their semantics. Legal reasoning with priorities over conflict-resolution principles (Prakken & Sartor, 1997) is the closest prior work on ordering several ways of resolving a conflict; FULCRUM orders criteria within a single comparison rather than choosing among resolution principles.

Gradual and neural argumentation. Gradual semantics (Baroni et al., 2018; Rago et al., 2016) and the neural reading of Potyka (2021) make argument strength numeric or learnable but yield one ranking; argumentative language models (Freedman et al., 2025) and LLM-ASPIC$^+$ (Fang et al., 2025) combine language models with argumentation under fixed orderings.

Preference learning and social choice. Scalar reward models (Bradley & Terry, 1952; Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023) share the single-ordering representation; non-transitive preference learning (Chen & Joachims, 2016; Balduzzi et al., 2018; Munos et al., 2024; Swamy et al., 2024; Wu et al., 2025) represents cycles with latent vectors or general pairwise probabilities. Zermelo's model of tournament outcomes (Zermelo, 1929) is the ancestor of Bradley–Terry; Kemeny and Slater optimal orderings (Kemeny, 1959; Slater, 1961) define the ceiling of Theorem 5, whose computation is NP-hard on tournaments (Alon, 2006; Charbit et al., 2007). That aggregating transitive criteria yields intransitive preference is classical (May, 1954; Tversky, 1969; Fishburn, 1991); Ganzer-Ripoll et al. (2019) aggregate across agents, whereas we vary the deciding criterion across pairs within one reasoner. HelpSteer2 and HelpSteer2-Preference (Wang et al., 2024b;a) provide the multi-aspect human preference data used in Section 6; PrefLib (Mattei & Walsh, 2013) and the Sushi data (Kamishima, 2003) provide the ranking data.