Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What do large language model panels add to multi-criteria decisions? Order-invariant q-rung dual hesitant preference aggregation and a hybrid supercapacitor case study

Omar S. Alhadithi
Tun Hussein Onn University of Malaysia
Lukman Audah
hanif@uthm.edu.my

Tun Hussein Onn University of Malaysia
Mohammed A. Alhartomi
University of Tabuk
Ameer Alhassan
Real Estate Bank
Mustafa Maad Hamdi
University of Anbar
Abdinasir Hirsi
Jamhuriya University of Science and Technology
Godwin Okon Ansa
Akwa Ibom State University
Abdullahi Farah
Somtel Telecommunication Company
Yazan Aljeroudi
Al-Farabi University

Research Article

Keywords: Large language models, q-Rung dual hesitant fuzzy preference relations, Order invariance, Multi-criteria decision-making, Uncertainty quantification, Hybrid supercapacitors

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11556583/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

What do large language model panels add to multi-criteria decisions? Order-invariant q-rung dual hesitant preference aggregation and a hybrid supercapacitor case study

Omar S. Alhadithi[1,2*], Lukman Audah[3*], Mohammed A. Alhartomi[4], Ameer Alhassan[5], Mustafa Maad Hamdi[6], Abdinasir Hirsi[7], Godwin Okon Ansa[8], Abdullahi Farah[9], Yazan Aljeroudi[10]

[1*]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, Johor, Malaysia.
[2]Department of Cybersecurity Engineering, Faculty of Engineering, Al-Karkh University of Science, Baghdad, Iraq.
[3*]Advanced Telecommunication Research Center, Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, Johor, Malaysia.
[4]Department of Electrical Engineering, University of Tabuk, Tabuk, 71491, Saudi Arabia.
[5]Information Security Department, Real Estate Bank, Baghdad, Iraq.
[6]Department of Computer Science, College of Computer Science and Information Technology, University of Anbar, Ramadi, 31001, Iraq.
[7]Faculty of Engineering, Jamhuriya University of Science and Technology, Mogadishu, 2602, Somalia.
[8]Department of Computer Science, Faculty of Computing and Information Technology, Akwa Ibom State University, Mkpat Enin, 520101, Akwa Ibom, Nigeria.
[9]Engineering Department, Somtel Telecommunication Company, Bosaso, 25290, Somalia.
[10]Department of Cybersecurity Engineering, Al-Farabi University, Baghdad, Iraq.

*Corresponding author(s). E-mail(s): eng.omars89@gmail.com;
hanif@uthm.edu.my;
Contributing authors: malhartomi@ut.edu.sa;
ameer.nadhum91@gmail.com; mustafa.maad.hamdi@uoanbar.edu.iq;
abdinasirhirsi@just.edu.so; godwinansa@aksu.edu.ng;
abdalla.gaash@somtelnetwork.net; yazan.aljeroudi@gmail.com;

Abstract

Large language models (LLMs) increasingly replace human experts in multi-criteria decision analysis, but their contribution to decisions computable from published data is rarely examined. This is studied with q-rung dual hesitant fuzzy preference relations elicited from five LLMs of five developers. The standard group relation, built with the weighted geometric operator and a reciprocity rebuild, depends on the order of the criteria; a q-power-mean aggregation with row-sum priority weights is proved to be order invariant, to preserve the q-rung constraint and to keep every weight positive, and without consensus adjustments the results depend on each hesitant set only through the means of the q-th powers of its degrees. In a hybrid supercapacitor selection problem built from verified full-cell data, the standard construction yields up to 26 rankings over the criterion orderings, the proposed construction one; the elicited judgments, however, shift with the order of the prompt (four permutation tests, $\boldsymbol{p \leq 0.009}$). Single central judgments per entry reproduce the hesitant-set rankings. With the decision table in the prompt, all 18 direct-data configurations reproduce the LLM-pipeline ranking; without it, the LLM performance judgments diverge from the evidence and change the ranking. Under ranking by multi-objective optimization by ratio analysis plus the full multiplicative form, resampling of models and sessions places the lithium-ion capacitor first in 0.748 (standard) and 1.000 (fixed-bound variant) of the draws; with literature source rows also resampled and the rebuilt table as performance matrix, the shares are 0.533 and 0.504. The LLM judgments are not validated against human experts.

Keywords: Large language models, q-Rung dual hesitant fuzzy preference relations, Order invariance, Multi-criteria decision-making, Uncertainty quantification, Hybrid supercapacitors

1 Introduction

Multi-criteria decision methods require two kinds of judgment: the relative importance of the criteria and the performance of each alternative under each criterion. When no panel of domain experts is available, large language models (LLMs) are increasingly used to supply both, for instance the pairwise comparisons and hierarchies of the analytic hierarchy process [1, 2], criterion weights [3] or the weights and decision matrices of complete analyses [4]. Three questions arise. The first is whether the aggregation of the judgments is itself free of artifacts: LLM judgments depend on the

order in which options are presented [5, 6], and an aggregation that depends on the order of the criteria adds a second, avoidable source of order dependence. The second is how much of the elaborate representation of the judgments enters the result. The third is what the LLM-elicited judgments add to a decision that could be computed from published data directly, and how uncertain the resulting ranking is once the disagreement between models and the uncertainty of the source data are propagated.

These questions are examined for q-rung dual hesitant fuzzy preference relations, which represent a judgment by sets of membership and non-membership degrees under a relaxed constraint [7, 8] and support consistency and consensus procedures for group decisions [9], combined with MULTIMOORA ranking [10]. The judgments are elicited from a panel of five LLMs from five developers. Aggregating sampled sessions of one model improves its reasoning accuracy [11], but such sessions are not independent judgments: repeated runs of one model can show near-zero variation [12], whereas the accuracy of LLM ensembles has been obtained by aggregating different models [13]. The demonstration case, the selection of hybrid supercapacitors for battery electric vehicles, has sparse evidence and conflicting criteria, and its published full-cell data allow the LLM-elicited performance judgments to be compared with the evidence they summarize.

The contributions are as follows.

1. The standard construction of the group relation, which combines the weighted geometric operator with a rebuild of the lower triangle and a crispification by the minimum score, is shown to depend on the order of the criteria. An aggregation by a q-power mean with row-sum priority weights is proved to preserve the q-rung constraint and reciprocity, to be invariant to the order of the criteria and of the panel members, and to bound every weight away from zero (Propositions 1–3). The invariance concerns the aggregation of judgments already elicited: a prompt-order experiment shows that the judgments themselves shift with the order of criteria and alternatives in the prompt, so that aggregation invariance does not imply invariance of the pipeline.
2. A sufficiency property is established: in the absence of consensus adjustments, projections and padding, the weights and rankings depend on each hesitant set only through the means of the q-th powers of its degrees (Proposition 4). A representation ablation and a control in which a single central judgment is elicited per entry show that, in this case, eliciting hesitant sets adds no decision information over a single central judgment, so that hesitant elicitation is optional; reducing elicited sets to one endpoint, by contrast, changes the results.
3. The contribution of the LLM panel to the decision is quantified. With the decision table in the prompt, the LLM-elicited performance judgments reproduce the table, and every direct-data configuration, including uniform weights, yields the same ranking as the LLM pipeline. Without the table, the panel changes the decision through its performance judgments, which then diverge from the published evidence. The criterion-importance judgments are stable across these conditions, but their usefulness is not validated against human judgments.
4. The uncertainty of the ranking is estimated by hierarchical resampling of models, sessions within models and literature source rows, with the decision table rebuilt

in every draw; independent perturbation of the elicited degrees is retained as a secondary, conditional sensitivity analysis. Standard and fixed-bound MULTIMOORA are treated as co-equal variants, and the dependence of the top rank on this choice is reported as a result.

The code, prompts, raw model responses and data are released with the article.

The scope of the claims is limited. The LLM-elicited judgments are not compared with those of human experts, and the agreement reported is agreement among models. The decision tables rest on few sources per cell, several of them on a single source, contain no cost criterion, and encode two ordinal criteria on a ratio scale. The rankings reported are therefore properties of the procedure applied to these data and not a technology recommendation.

Sections 2–8 present related work, methods, formal properties, the case study, results, discussion and conclusions; Online Resource 1 contains the prompts, proofs and extended tables; sections, tables, figures and algorithms numbered with the prefix S refer to it.

2 Related work

q-rung and dual hesitant preference relations

q-rung orthopair fuzzy sets relax the constraint of intuitionistic and Pythagorean fuzzy sets to $\mu^q + \nu^q \leq 1$ [7]; weighted averaging and geometric operators were introduced for them [14]. q-rung dual hesitant fuzzy sets add sets of possible degrees and were aggregated with Heronian-mean operators [8], and the corresponding preference relations were embedded in the analytic hierarchy process with consistency and consensus procedures [9]. Subsequent operators for q-rung dual hesitant information include extended power averages based on Archimedean operations [15], aggregation with entropy-based attribute weights [16] and Frank-norm power partitioned Heronian means [17]; consistency and consensus models have also been developed for linguistic interval-valued q-rung orthopair fuzzy preference relations [18]. These relations generalize the additive reciprocal fuzzy preference relations of group decision making [19], whose priority vectors are commonly derived from row sums [20]. For operators that are not self-dual under the reciprocal complement, a group relation rebuilt from one triangle depends on the order of the criteria (Remark 1).

FWZIC, MULTIMOORA and rank reversal

The fuzzy-weighted zero-inconsistency (FWZIC) method derives criterion weights from expert scores of the criteria, without pairwise comparisons and therefore without a consistency test [21]; it has been extended to spherical [22], q-rung orthopair 2-tuple linguistic [23], q-rung orthopair rough [24, 25] and q-rung orthopair probabilistic hesitant [26] environments and combined with ranking methods such as the fuzzy decision by opinion score method, CODAS, CoCoSo, GRA–TOPSIS or MULTIMOORA. MULTIMOORA combines a ratio system, a reference point and a full multiplicative form through dominance [10]; its fuzzy extensions have been surveyed [27] and applied, for instance, to energy-storage technologies [28], the location of electric-vehicle charging stations under Fermatean fuzzy information [29], waste disposal [30] and, with

dual hesitant fuzzy sets and the best–worst method, the selection of unmanned aerial vehicles [31]. For hybrid supercapacitors in electric vehicles, FWZIC weights have been combined with fuzzy MULTIMOORA on four capacitor chemistries and three criteria, with LLM-generated synthetic preference panels used as a robustness experiment [32]; the present study replaces that construction by q-RDHF preference relations elicited from a panel of LLMs, an order-invariant aggregation, an enlarged literature-verified decision problem and the controls described in Section 3.6. The dominance-theory aggregation of the three components requires multiple comparisons and can involve circular reasoning, and has been replaced, for example, by an aggregation based on half-quadratic theory [33]. Rank reversal under the addition, removal or replacement of alternatives is documented for several methods [34, 35], and simulation studies relate its probability to the method and to the rank of the deleted alternative [36]; a primary analysis specific to MULTIMOORA was not located, so that general tests are applied here.

LLMs as sources of judgments

LLMs have supplied pairwise comparisons for the analytic hierarchy process under criteria they generated [1], constructed analytic hierarchy models after fine-tuning on domain documents [2], assigned criterion weights as a multi-agent system with outlier removal and an iterative consensus based on Kendall's W [3], derived fuzzy preference relations from the debate comments of human experts [37] and quantified communicative traits of participants for large-scale group decision making [38]. In a comparison with domain experts within one decision structure, the criterion weights generated by five LLMs agreed closely with those of the experts, whereas the decision matrices and rankings differed more [4]. LLM judges show position, verbosity and self-enhancement biases [39], and their preferences can change with the order of presentation alone [5]. Position bias varies across judges and tasks and is not attributable to chance [40], order effects can lead models to select strictly inferior options [6], and aggregation over a full factorial design of item orders and labels removes response-order and label bias [41]. The scores of one LLM judge also vary across runs [42], and verbalized confidence tends to be overconfident [43, 44], so that hesitant degrees represent stated vagueness rather than calibrated uncertainty.

Positioning

Table 1 compares the closest prior work with this study. Cells are filled from the full texts of the studies, except for [1, 37], for which the title and abstract were consulted; a question mark marks information that could not be determined from that text and does not indicate that a study lacks the property. Two of the four LLM-based studies elicit judgments from several models and one from a single model per run. The full texts of [3, 4] represent the elicited judgments by crisp values rather than by hesitant or q-rung sets; the abstracts of [1, 37] describe crisp pairwise comparisons and fuzzy preference relations, respectively, and do not mention hesitant or q-rung representations, so that their use in the full texts of these two studies is not excluded. None of the eleven studies states order invariance of the weights in the consulted text. In the seven conventional studies, the judgments are supplied by three or four domain

experts, by an illustrative example or by published data, and expert judgments are combined by aggregation operators or rank-based group scoring without a consensus-reaching process, with the exception of the consistency and consensus procedures for q-RDHF preference relations [9].

3 Methods

The pipeline (Fig. 1) builds a decision table from published data, elicits hesitant judgments from a panel of LLMs, derives weights from the elicited preference relations, ranks the alternatives with two MULTIMOORA variants and quantifies the ranking uncertainty; control conditions replace parts of the pipeline.

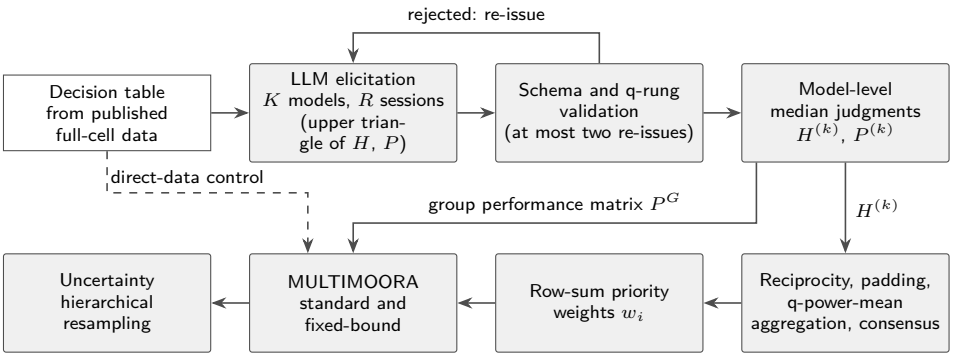


Fig. 1 Pipeline. The table is supplied in the prompt; the elicited preference relations determine the weights and the elicited performance judgments enter the ranking. In the direct-data control (dashed), the table replaces the elicited performance judgments.

3.1 Problem formulation and preliminaries

Let $A = \{A_1, \ldots, A_m\}$ be the alternatives and $C = \{C_1, \ldots, C_n\}$ the criteria, partitioned into benefit criteria J^+ and cost criteria J^-; in both problems of this study $J^- = \emptyset$, and the general form is retained because each criterion direction is applied exactly once. Each of K panel members (LLMs) supplies a preference relation $H^{(k)} = (h_{ij}^{(k)})_{n\times n}$, in which $h_{ij}^{(k)}$ expresses the degree to which C_i is more important than C_j, and a performance matrix $P^{(k)} = (p_{aj}^{(k)})_{m\times n}$. The outputs are weights $\mathbf{w}$ with $\sum_i w_i = 1$, three component rankings and an integrated ranking, and the distribution of the ranking under resampling.

Definition 1 (q-rung dual hesitant fuzzy element) For an integer $q \geq 1$, a q-rung dual hesitant fuzzy element (q-RDHFE) is a pair $h = \langle M, N\rangle$ of finite tuples of membership degrees $M = (\mu^{(1)}, \ldots, \mu^{(|M|)})$ and non-membership degrees $N = (\nu^{(1)}, \ldots, \nu^{(|N|)})$ in $[0, 1]$ with [7, 8]

$$(\max M)^q + (\max N)^q \leq 1. \tag{1}$$

Table 1 Comparison of the closest prior work with the proposed framework.

Study	Representation	Criteria weighting	Ranking	Elicitation source	Consensus mechanism	Order invariance stated
[8]	q-RDHF elements	Given weight vector (numerical example)	q-RDHF Heronian-mean aggregation	Illustrative supplier-selection example adopted from the literature	— (single decision matrix)	Not stated
[9]	q-RDHF preference relations	AHP priority weights (new calculation method)	AHP	Three experts (illustrative hospital-infection risk case); expert weights from a decision-maker preference relation	Consistency and consensus checking and improving	Not stated
[23]	q-ROF 2-tuple linguistic	q-ROF2TL-FWZIC	q-ROF2TL-CODAS	Three domain experts (criterion importance); evaluation data of 50 applications adopted from a prior study	Averaging of expert scores within FWZIC; no consensus-reaching process	Not stated
[24]	q-ROF rough sets	q-ROFRS–FWZIC	q-ROFRS–FDOSM	Three domain experts (criterion importance and opinion matrices)	Aggregation of expert judgments and rank-based group scoring; no consensus-reaching process	Not stated
[26]	q-ROF probabilistic hesitant	q-ROPHFS–FWZIC	q-ROPHFS–FDOSM scoring + MULTIMOORA	Three domain experts (criterion importance and opinion matrices)	Einstein arithmetic-mean aggregation and arithmetic mean of expert scores; no consensus-reaching process	Not stated
[30]	q-ROF	Entropy and discrimination (objective) + operator/score (subjective)	q-ROF MULTIMOORA	Four domain experts (municipal, environmental and sustainable-development sectors)	q-ROF weighted averaging with computed expert weights; no consensus-reaching process	Not stated
[28]	Intuitionistic fuzzy + crisp	Four assumed weighting scenarios (equal and dimension-priority)	MULTIMOORA-IFN2	Published technology data combined from several sources	— (no group)	Not stated
[1]	Crisp pairwise comparisons	AHP (LLM-generated criteria)	AHP	One LLM per run (GPT-3.5, GPT-4)	—	Not stated
[37]	Fuzzy preference relations	— (expert weights from participation)	Group decision-making selection	LLM sentiment analysis of human debate comments	?	Not stated
[3]	Weights extracted from LLM outputs	LLM multi-agent weighting (WALMAS)	Weighted-sum overlay (landfill suitability map)	Several LLMs (GPT-4o, Gemini, Copilot) as agents with specialist roles	IQR outlier removal and iterative negotiation based on Kendall's W	Not stated
[4]	Crisp values from a 9-point linguistic scale	RANCOM	MARCOS; aggregation by compromise fuzzy ranking	Domain experts and five LLMs; repeated LLM evaluations	Group consensus ranking by compromise fuzzy ranking	Not stated
This study	q-RDHF preference relations	Row-sum priority weights of the q-power-mean group relation (Prop. 3)	Standard and fixed-bound q-RDHF MULTIMOORA; hierarchical resampling	Panel of five LLMs from five developers; direct-data and knowledge-only controls	q-power-mean aggregation and consensus fusion with projection (Prop. 2)	Aggregation level (Props. 2, 3)

Notes. "?": not determinable from the consulted text, not evidence of absence; "Not stated": not stated in the consulted text; "—": not applicable (single judge, no group). Abbreviations: q-ROF, q-rung orthopair fuzzy; q-RDHF, q-rung dual hesitant fuzzy; q-ROFRS, q-ROF rough set; q-ROPHFS, q-ROF probabilistic hesitant fuzzy set; q-ROF2TL, q-ROF 2-tuple linguistic; FWZIC, fuzzy-weighted zero-inconsistency; FDOSM, fuzzy decision by opinion score method; CODAS, combinative distance-based assessment; RANCOM, ranking comparison; MARCOS, measurement of alternatives and ranking according to compromise solution; IQR, interquartile range.

The value $q = 3$ is used throughout. On construction, every degree is clipped to $[0, 1]$, the degrees of each tuple are sorted in non-decreasing order, and a pair that violates (1) is mapped onto the boundary by the radial projection

$$\Pi\left(\langle M, N\rangle\right) = \langle \kappa M, \kappa N\rangle, \qquad \kappa = \min\left\{1,\ \left[(\max M)^q + (\max N)^q\right]^{-1/q}\right\}. \tag{2}$$

The score and the accuracy of h are

$$S(h) = \frac{1}{|M|}\sum_l \left(\mu^{(l)}\right)^q - \frac{1}{|N|}\sum_l \left(\nu^{(l)}\right)^q \in [-1, 1], \tag{3}$$

$$F(h) = \frac{1}{|M|}\sum_l \left(\mu^{(l)}\right)^q + \frac{1}{|N|}\sum_l \left(\nu^{(l)}\right)^q \in [0, 1], \tag{4}$$

and its reciprocal complement is

$$h^c = \langle \mathrm{rev}(N), \mathrm{rev}(M)\rangle, \tag{5}$$

where rev reverses a tuple, so that $S(h^c) = -S(h)$. A q-RDHF preference relation $H = (h_{ij})$ is reciprocal if [9]

$$h_{ii} = \langle (0.5), (0.5)\rangle, \qquad h_{ji} = h_{ij}^c \quad (1 \le i < j \le n); \tag{6}$$

every relation formed in the pipeline is made reciprocal by setting the diagonal and rebuilding the lower triangle from the upper triangle.

3.2 Decision tables

The evidence base consists of peer-reviewed full-cell reports (2015–2026) with verified identifiers; no value is estimated or read from a graph. A device class is admitted if it pairs a Faradaic with a capacitive electrode, has full-cell reports from at least three groups and a source discussing vehicle use. Energy and power density are taken on the total active mass of both electrodes; values on one electrode or without a stated basis are excluded. Cycle life is the number of cycles in tests that end with at least 80 % capacity retention; because the tests end at different cycle numbers, current densities and retentions, it is a lower bound rather than a measured end of life. Safety is an electrolyte flammability ordinal (aqueous 3, ionic liquid 2, organic carbonate or glyme 1; a gel takes the class of its solvent), following the ordering of the electrolyte literature [45–47], with the modal class of the sources of an alternative; technology readiness is the level on the technology readiness scale introduced by NASA [48], in the form adopted for European research and innovation funding in the 2014–2020 programming period (Horizon 2020) [49], assigned from the most advanced demonstration. The safety definition shown to the models read “ionic-liquid/gel = 2”; no alternative has a gel as its modal class, so no table value is affected (Section S1).

Each cell is the median of the admissible values, with the number of distinct sources recorded and the identifiers listed in Table S19. Every quantity is normalized

by the ratio to the best alternative, $\tilde{x}_{aj} = x_{aj} / \max_b x_{bj}$; in the reduced problem, Performance and Durability are geometric means of the normalized energy and power density and of cycle life and capacity retention. The ratio normalization treats the ordinal scales as ratio scales; alternative encodings are examined in Section 3.6.

3.3 Elicitation of LLM judgments

Panel and sessions

The panel comprises one current flagship model from each of five developers, accessed through one API gateway (Table 3); a sixth model serves as an anchor for comparison only. Each model is queried in $R = 5$ independent single-message sessions without system prompt or persona, at temperature 0.7 where accepted and the provider default otherwise; identifier, provider, temperature, cost and prompt hash are recorded for every call.

Prompt and validation

The prompt (Section S1) contains the problem, the criterion definitions, the normalized table and the medians with source counts, and requests for every pair (C_i, C_j), $i < j$, and every alternative and criterion at least two membership and two non-membership degrees satisfying (1), in a fixed JSON schema. A response is accepted only if it parses, contains every judgment with at least two degrees per tuple in $[0, 1]$, and satisfies (1); a rejected session is re-issued with the identical prompt at most twice, and every attempt is stored.

Model-level judgments

For model k with R_k accepted sessions, each entry of $H^{(k)}$ and $P^{(k)}$ retains the median over sessions of the per-session minimum and of the per-session maximum of each tuple,

$$\begin{aligned} M^{(k)} &= \operatorname{sort}\left(\operatorname{med}_r \min M_r,\ \operatorname{med}_r \max M_r\right), \\ N^{(k)} &= \operatorname{sort}\left(\operatorname{med}_r \min N_r,\ \operatorname{med}_r \max N_r\right), \end{aligned} \tag{7}$$

followed by Π; every model-level entry therefore carries two degrees per tuple. Agreement is measured by rank correlations and Kendall's coefficient of concordance W of the model-wise weights and rankings, and by the between-model share $\beta = V_{\text{between}}/(V_{\text{between}} + V_{\text{within}})$ of the score variance, averaged over the judgments of H and of P.

Control conditions of the elicitation

Three conditions repeat the elicitation with the same models, settings and validation. In the knowledge-only condition, the decision table and the medians are removed from the prompt and the models are instructed to judge from their own knowledge; the criterion definitions are retained. In the prompt-order condition, the criteria and alternatives are listed in three randomized orders (in one order of the reduced problem only the alternatives), with three sessions per order and model; the normalized table follows the permuted order, whereas the reference table of raw medians keeps its

column order, and the elicited relations are mapped back to a canonical orientation. The between-order share of the score variance is tested against random permutation of the sessions across orders (2000 permutations). In the single-judgment condition, the prompt requests exactly one membership and one non-membership degree per entry. All prompts are reproduced in Section S1.

3.4 Row-sum priority weighting of q-RDHF preference relations

The weighting stage aggregates the model-level relations into a group relation and derives the weights from its scores (Algorithm 1).

Relation to FWZIC

The stage shares with the fuzzy-weighted zero-inconsistency method [21] the derivation of weights without a pairwise-comparison consistency ratio. It differs in its input and its weight rule: FWZIC weights criteria from direct importance scores, whereas the present stage uses pairwise preference relations and normalized row sums. Reciprocity of the relations is enforced by construction; it does not imply, and the stage does not claim, consistency of the preferences.

Padding and aggregation

Tuples are aligned to a common length L by appending to M copies of

$$\bar{\mu}_{\mathrm{pad}} = (\max M)^{\zeta} (\min M)^{1-\zeta}, \qquad \zeta = 0.5, \tag{8}$$

and likewise to N; padding is applied within each relation and then across relations. In the panel data every off-diagonal entry has two degrees per tuple, so that only the diagonal is padded.

Definition 2 (q-power mean) For q-RDHFEs $h_1, \ldots, h_K$ of common length L and weights $\omega_k \geq 0$, $\sum_k \omega_k = 1$,

$$\mathrm{QPM}_{\omega}(h_1, \ldots, h_K) = \left\langle \left(\left(\sum_k \omega_k (\mu_k^{(l)})^q \right)^{1/q} \right)_{l=1}^{L}, \left(\left(\sum_k \omega_k (\nu_k^{(l)})^q \right)^{1/q} \right)_{l=1}^{L} \right\rangle. \tag{9}$$

With $\omega_k = 1/K$, the group relation is

$$G^{(0)} = \left(\mathrm{QPM}_{\omega} \left(h_{ij}^{(1)}, \ldots, h_{ij}^{(K)} \right) \right)_{n \times n}. \tag{10}$$

Consensus

The distance between two aligned relations is

$$d(H, G) = \frac{1}{2Ln(n-1)} \sum_{i \neq j} \sum_{l=1}^{L} \left(\left| \left(\mu_{H,ij}^{(l)}\right)^q - \left(\mu_{G,ij}^{(l)}\right)^q \right| + \left| \left(\nu_{H,ij}^{(l)}\right)^q - \left(\nu_{G,ij}^{(l)}\right)^q \right| \right). \tag{11}$$

A relation with $d(H^{(k)}, G^{(0)}) \geq \tau$ is moved towards the fixed target $G^{(0)}$ by applying the fusion rule [9]

$$\phi_\sigma(x, y) = \frac{x^\sigma y^{1-\sigma}}{x^\sigma y^{1-\sigma} + (1-x)^\sigma (1-y)^{1-\sigma}} \tag{12}$$

to every degree x of $h_{ij}^{(k)}$ and the degree y of $g_{ij}^{(0)}$ in the same layer ($\phi_\sigma = 0.5$ for a vanishing denominator), sorting, projecting by Π and rebuilding reciprocity. The projection is required because (12) can leave the q-rung domain. Fusion stops when $d < \tau$ or after $T_{\max}$ steps, with $\tau = 0.1$, $\sigma = 0.5$ and $T_{\max} = 50$; the adjusted relations are aggregated again by (10) into the final group relation $G = (g_{ij})$.

Weights

With $r_{ij} = (1 + S(g_{ij}))/2$, the weights are the normalized row sums

$$w_i = \frac{\sum_j r_{ij}}{\sum_i \sum_j r_{ij}}, \qquad i = 1, \ldots, n, \tag{13}$$

without a shift by the minimum score.

3.5 Ranking: standard and fixed-bound MULTIMOORA

The group performance matrix is

$$p_{aj}^G = \mathrm{QPM}_\omega \left(p_{aj}^{(1)}, \ldots, p_{aj}^{(K)}\right), \tag{14}$$

with scores $s_{aj} = S(p_{aj}^G)$ and accuracies $f_{aj} = F(p_{aj}^G)$. Two variants of MULTIMOORA [10, 27] are applied and reported as co-equal; they differ only in the transformation of the scores.

Standard variant

The scores are normalized by the column norm, $n_{aj} = s_{aj}/(\|s_{\cdot j}\|_2 + \varepsilon)$, and weighted, $v_{aj} = w_j n_{aj}$, with $\varepsilon = 10^{-12}$. The ratio system ranks by $Y_a = \sum_{J^+} v_{aj} - \sum_{J^-} v_{aj}$ and the reference point by the Tchebycheff distance $Z_a = \max_j |v_{aj} - \rho_j|$ to $\rho_j = \max_a v_{aj}$ ($j \in J^+$) or $\min_a v_{aj}$ ($j \in J^-$). Because the normalized scores can be negative, the full multiplicative form requires a positive transformation; the implementation rescales each weighted column to $[0, 1]$,

$$\tilde{v}_{aj} = \frac{v_{aj} - \min_b v_{bj}}{\max_b v_{bj} - \min_b v_{bj} + \varepsilon}, \qquad U_a = \frac{\prod_{J^+} (\tilde{v}_{aj} + \varepsilon)^{w_j}}{\prod_{J^-} (\tilde{v}_{aj} + \varepsilon)^{w_j} + \varepsilon}. \tag{15}$$

This rescaling assigns $\tilde{v}_{aj} = 0$ to the worst alternative on each criterion, so that the order among alternatives that carry such a zero factor is determined by ε^{w_j} rather than by the data; the sensitivity of the component and of the integrated ranking to ε is reported in Section 6.5.

Fixed-bound variant

The scores are mapped to $\hat{n}_{aj} = (s_{aj} + 1)/2 \in [0, 1]$ using the fixed range of S. The components are $\hat{Y}_a = \sum_{J^+} w_j \hat{n}_{aj} - \sum_{J^-} w_j \hat{n}_{aj}$, $\hat{Z}_a = \max_j w_j |\hat{n}_{aj} - \hat{\rho}_j|$ with the fixed ideal $\hat{\rho}_j = 1$ ($j \in J^+$) or 0 ($j \in J^-$), and $\hat{U}_a = \prod_{J^+} \hat{n}_{aj}^{w_j} / \prod_{J^-} \hat{n}_{aj}^{w_j}$. Since $S > -1$ for every element of the panel data, no factor vanishes and $\hat{U}_a$ does not depend on ε. Each component score depends only on the row of the alternative concerned.

Integration

In both variants, with $\pi_c(a)$ the position of A_a in component $c \in \{\mathrm{RS}, \mathrm{RP}, \mathrm{FMF}\}$, A_a dominates A_b if $|\{c : \pi_c(a) < \pi_c(b)\}| \geq 2$, and the integrated ranking orders the alternatives lexicographically by

$$\Big(-D_a,\ \tfrac{1}{3} \textstyle\sum_c \pi_c(a),\ -\sum_j f_{aj} \Big), \tag{16}$$

where D_a is the number of alternatives dominated by A_a.

3.6 Controls and uncertainty analysis

Direct-data baseline

The LLM-elicited performance matrix is replaced by the table itself, mapped to $\langle (x^{1/q}), ((1-x)^{1/q}) \rangle$ (score $2x - 1$), to $\langle (x), (1-x) \rangle$ or used crisp, with uniform, LLM or knowledge-only weights under both variants.

Representation ablation

Every element is replaced by its q-power-mean element (Section 4), by its smallest or largest degrees, by intuitionistic forms, or by crisp scores (Section S5).

Data-layer sensitivity

The table is rebuilt from the source rows (i) under four encodings of each ordinal criterion, (ii) with each source of each cell left out in turn, and (iii) with a bootstrap of the rows within each alternative and criterion ($B = 2000$), in which single-source cells keep their values.

Hierarchical resampling

The principal uncertainty estimate resamples, in each of $B = 2000$ draws, five panel members with replacement, for each drawn member its accepted sessions with replacement, and the source rows of the decision table as in (iii). Weights and performance matrix are rebuilt from the drawn sessions, and the ranking is computed (i) with the resampled LLM performance matrix and (ii) with the rebuilt table mapped directly, under both variants.

Conditional perturbation and rank reversal

As a secondary check, every elicited degree is multiplied by $1 + \delta$, δ uniform on $[-a, a]$ or truncated Gaussian with standard deviation $a/2$, $a \in \{0.05, 0.10\}$ (10 000 draws

per setting); the results are conditional on these levels, which do not represent an estimated error distribution. Rank reversal is tested with fixed weights by removing each alternative, adding a dominated alternative [34] and replacing a non-optimal alternative by a degraded one [35] (Section S7).

3.7 Use of large language models and AI-assisted tools

Large language models are the object of study: the models of Table 3 were queried through their programming interfaces under the protocol of Section 3.3, and their responses constitute the elicited judgments analyzed in this article; model identifiers, dates, prompt hashes and all raw responses are recorded in the data deposit. In addition, an AI-assisted tool (Claude, Anthropic) was used for code development, data processing and language editing of the manuscript; all outputs were reviewed by the authors, who take full responsibility for the content.

4 Formal properties

This section states four properties of the weighting and ranking stages; full proofs are given in Section S4. The propositions concern the aggregation of elicited judgments; they do not imply that the judgments themselves are invariant to the order in which a prompt lists the criteria. Elements are aligned by the padding rule (8); padding commutes with the complement for entries whose tuples have equal length, which holds in the panel pipeline (only diagonal entries are padded). For a permutation ψ of the criteria, $H^{\psi} = (h_{\psi(i)\psi(j)})$.

Proposition 1 (Closure and commutation with the complement) *Let $h_1, \dots, h_K$ be q-RDHFEs aligned to a common length L, and let $\omega_k \geq 0$ with $\sum_k \omega_k = 1$. Then:*

(a) if every h_k satisfies (1), then $\mathrm{QPM}_{\omega}(h_1, \dots, h_K)$ satisfies (1);
(b) $\mathrm{QPM}_{\omega}(h_1^c, \dots, h_K^c) = \mathrm{QPM}_{\omega}(h_1, \dots, h_K)^c$;
(c) $\Pi(h^c) = \Pi(h)^c$ for every pair $h = \langle M, N \rangle$ of tuples in $[0, 1]$.

Proof sketch (a) The q-th power of every aggregated degree is a convex combination of the q-th powers of the input degrees, so that the maxima satisfy (1). (b) The q-power mean treats membership and non-membership degrees identically, layer by layer, and therefore commutes with their exchange and reversal. (c) The factor κ in (2) is symmetric in M and N. The full proof is given in Section S4. □

Proposition 2 (Reciprocity and criterion-order invariance of the group relation) *Let $H^{(1)}, \dots, H^{(K)}$ be reciprocal q-RDHFPRs, let G be their entry-wise q-power mean (10), and let ψ be a permutation of the criteria. Then:*

(i) G is reciprocal; in particular, rebuilding its lower triangle from its upper triangle leaves it unchanged;
(ii) $\mathrm{QPM}_{\omega}\left((H^{(1)})^{\psi}, \dots, (H^{(K)})^{\psi}\right) = G^{\psi}$;

(iii) *(i) and (ii) continue to hold after any number of fusion steps (12), each followed by the projection (2) and the reciprocity rebuild, and every model-level relation undergoes the same number of fusion steps under every ordering of the criteria.*

Proof sketch Apply Proposition 1(b) entry-wise; fusion and the distance (11) treat all degrees and pairs alike (Section S4). □

Proposition 3 (Equivariance and strictly positive bounds of the weights) *Let G be a reciprocal q-RDHFPR on n criteria, $r_{ij} = (1 + S(g_{ij}))/2$, and w_i as in (13). Then:*

(a) *$R = (r_{ij})$ is an additive reciprocal fuzzy preference relation, that is, $r_{ij} + r_{ji} = 1$ and $r_{ii} = 1/2$ [19];*

(b) *$w_i = \frac{2}{n^2} \sum_j r_{ij}$, and*

$$\frac{1}{n^2} \le w_i \le \frac{2n-1}{n^2}, \qquad i = 1, \dots, n, \tag{17}$$

so that no criterion receives a zero weight;

(c) *$\mathbf{w}(G^\psi) = \mathbf{w}(G) \circ \psi$ for every permutation ψ.*

Proof sketch $S(g^c) = -S(g)$ gives $r_{ij} + r_{ji} = 1$; hence $\sum_{i,j} r_{ij} = n^2/2$ for every input, and each row sum lies between $1/2$ and $n - 1/2$ (Section S4). □

Remark 1 (Order dependence of the standard construction) A common construction forms the group relation with the q-rung orthopair weighted geometric operator, rebuilds the lower triangle from the upper triangle, and crispifies the weights by a shift by the minimum score. Layer-wise, the weighted geometric (WG) and weighted averaging (WA) operators [14] are

$$\mathrm{WG}_\omega(h_1, \dots, h_K) = \left\langle \prod_k (\mu_k)^{\omega_k}, \ \left(1 - \prod_k \left(1 - \nu_k^q\right)^{\omega_k}\right)^{1/q} \right\rangle, \tag{18}$$

$$\mathrm{WA}_\omega(h_1, \dots, h_K) = \left\langle \left(1 - \prod_k \left(1 - \mu_k^q\right)^{\omega_k}\right)^{1/q}, \ \prod_k (\nu_k)^{\omega_k} \right\rangle. \tag{19}$$

Exchanging the roles of μ and ν in (18) yields (19), so that $\mathrm{WA}_\omega(h_1^c, \dots, h_K^c) = \mathrm{WG}_\omega(h_1, \dots, h_K)^c$. Under this construction the pair (C_i, C_j) with $i < j$ is aggregated by WG, and the rebuilt entry $g_{ji} = g_{ij}^c = \mathrm{WA}_\omega\big((h_{ji}^{(k)})_k\big)$ is, in effect, aggregated by WA. Because WG and WA differ unless the aggregated judgments coincide, the group relation and every quantity derived from it depend on which criterion is listed first. Further, the crispification $\hat{w}_i \propto S(w_i) - \min_j S(w_j) + \varepsilon$ assigns the weight $\varepsilon / \sum_t \big(S(w_t) - \min_j S(w_j) + \varepsilon\big)$ to the criterion with the lowest score for every input, so that one weight is zero up to ε by construction. The q-power mean (9) and the row-sum weights (13) remove both effects by Propositions 2 and 3.

The last property concerns the information that the hesitant sets carry into the results. For an element $h = \langle M, N \rangle$, let

$$m_\mu(h) = \frac{1}{|M|} \sum_{l=1}^{|M|} \left(\mu^{(l)}\right)^q, \qquad m_\nu(h) = \frac{1}{|N|} \sum_{l=1}^{|N|} \left(\nu^{(l)}\right)^q \tag{20}$$

denote the means of the q-th powers of its degrees, so that $S(h) = m_\mu(h) - m_\nu(h)$ and $F(h) = m_\mu(h) + m_\nu(h)$ by (3)–(4).

Proposition 4 (Score sufficiency) *Consider the model-level relations $H^{(1)}, \ldots, H^{(K)}$ and the model-level performance matrices $P^{(1)}, \ldots, P^{(K)}$ that enter the weighting and ranking stages, and assume that (i) every input element satisfies (1), so that the projection Π is not applied to it; (ii) all off-diagonal elements of the relations, and all elements of the performance matrices, have tuples of one common length L, so that padding alters only the diagonal elements $\langle (0.5), (0.5) \rangle$; and (iii) no model-level relation undergoes a consensus-fusion step. Then the weights $\mathbf{w}$, the score matrix (s_{aj}) and the accuracy matrix (f_{aj}) of the ranking stage, and hence all component rankings and the integrated ranking, depend on the input elements only through the pairs (m_μ, m_ν) of (20).*

Proof sketch The q-power mean is linear in (m_μ, m_ν); complement and diagonal padding preserve the pair; without fusion $G = G^{(0)}$ (Section S4). □

Under the assumptions of Proposition 4, replacing every element by its q-power-mean element $\langle (m_\mu^{1/q}), (m_\nu^{1/q}) \rangle$ leaves the results unchanged, whereas replacing it by an endpoint of its tuples does not; the role of each assumption is discussed in Section S4.

5 Case study

Metal-ion hybrid capacitors pair a battery-type with a capacitive electrode in order to combine the energy density of batteries with the power density and cycle life of supercapacitors [50–52]. Four device classes were retained as alternatives from the outset: Li-ion, Na-ion and K-ion capacitors and the Zn-ion hybrid supercapacitor. Of four further candidates, the dual-ion hybrid capacitor and the Mg-ion hybrid supercapacitor were excluded because no peer-reviewed source discussing vehicle or transport use of either class was located; the aqueous Ni-based asymmetric hybrid capacitor and the lead–carbon hybrid capacitor passed the screen on prospective statements, and the lead–carbon device was then excluded because no energy or power density on the two-electrode basis was available. The extraction comprises 153 value rows from 50 sources.

Two problems are built (Table 2). The enlarged problem, with five alternatives and five criteria, is the main case; the reduced problem, with four alternatives and three pillars, permits the enumeration of all criterion orderings with little computation. Cost is not a criterion, since a published cost estimate was located only for the excluded lead–carbon device. The energy and power densities of the Na-ion capacitor, the Zn-ion

device and the Ni-based device rest on a single source each; the single Zn-ion value was obtained with an MXene anode. The technology readiness level is a single assignment per alternative, and only the Li-ion capacitor has an in-vehicle demonstration, in an auxiliary application. The panel and the elicitation statistics are listed in Table 3.

Table 2 Decision tables. Left: medians of the published full-cell values with the number of independent sources (†: single source); E, P: energy and power density on the total active mass of both electrodes; N_{80}: cycles in tests ending with at least 80 % capacity retention, from tests with differing end points (a lower bound); Ret.: capacity retention; Saf.: electrolyte flammability ordinal. Right: normalized criteria (ratio to the best alternative) as supplied to the models. Source identifiers per cell: Table S19.

(a) Reduced problem

	Median of published values (n sources)					Normalized criteria		
Alternative	E (Wh kg^{-1})	P (W kg^{-1})	N_{80}	Ret. (%)	Saf.	Performance	Safety	Durability
Li-ion	172 (3)	11000 (3)	6000 (4)	85.1 (5)	1	0.9955	0.3333	0.7456
Na-ion	106.5 (1)†	11100 (1)†	6500 (2)	76.25 (4)	1	0.7869	0.3333	0.7346
K-ion	100 (3)	8109 (2)	6000 (7)	87.4 (8)	1	0.6517	0.3333	0.7556
Zn-ion	98.6 (1)†	2481 (1)†	10000 (8)	91.85 (8)	3	0.3580	1.0000	1.0000

(b) Enlarged problem

	Median of published values (n sources)					Normalized criteria				
Alternative	E (Wh kg^{-1})	P (W kg^{-1})	N_{80}	Saf.	TRL	Energy	Power	Cycle life	Safety	TRL
Li-ion	172 (3)	11000 (3)	6000 (4)	1	7	1.0000	0.9910	0.6000	0.3333	1.0000
Na-ion	106.5 (1)†	11100 (1)†	6500 (2)	1	4	0.6192	1.0000	0.6500	0.3333	0.5714
K-ion	100 (3)	8109 (2)	6000 (7)	1	4	0.5814	0.7305	0.6000	0.3333	0.5714
Zn-ion	98.6 (1)†	2481 (1)†	10000 (8)	3	3	0.5733	0.2235	1.0000	1.0000	0.4286
Ni-aq.	23.45 (1)†	4598 (1)†	6000 (4)	3	4	0.1363	0.4142	0.6000	1.0000	0.5714

Table 3 LLM panel, sampling temperature ("default": parameter not accepted, provider default), accepted sessions over calls including re-issues, and cost as reported by the gateway, for the main elicitation.

				Accepted / calls		Cost (USD)	
Role	Developer	Model identifier	Temperature	4 × 3	5 × 5	4 × 3	5 × 5
Panel	OpenAI	`openai/gpt-6-astra`	default	5/5	5/5	0.253	0.557
Panel	Anthropic	`anthropic/claude-fable-5.1`	default	5/5	5/5	0.517	0.967
Panel	Google	`google/gemini-3.1-pro-preview`	0.7	5/5	5/5	0.250	0.343
Panel	Mistral AI	`mistralai/mistral-large-4-0`	0.7	5/5	5/5	0.081	0.110
Panel	Meta	`meta-llama/llama-4-maverick`	0.7	5/5	3/10	0.002	0.009
Anchor	OpenAI	`openai/gpt-4o`	0.7	5/6	3/10	0.045	0.146
Total						1.148	2.131

6 Results

All results use $q = 3$, $\tau = 0.1$, $\sigma = 0.5$ and $T_{\max} = 50$; extended tables are given in Online Resource 1.

6.1 Order dependence of the standard construction

The weights and rankings were recomputed for all 120 orderings of the criteria of the enlarged and all 6 of the reduced problem, for $q = 1, \ldots, 5$, on the same elicited judgments (Table 4, Fig. S7). Under the standard construction of Remark 1, the largest spread of a weight over the orderings is 0.232–0.787 in the enlarged and 0.501–0.981 in the reduced problem, 19 to 26 distinct integrated rankings occur in the enlarged and 2 to 3 in the reduced problem, and the lower bound of Proposition 3 is violated for every q because one weight is zero up to ε. Under the q-power mean with row-sum weights, the spread is zero and a single ranking results for every q; the results are also invariant to all 120 orderings of the panel members. These statements concern the aggregation of fixed judgments; prompt-order effects are reported in Section 6.6.

Table 4 Criterion-order dependence on the same elicited judgments: largest spread of a weight over all orderings of the criteria and number of distinct integrated rankings, for the standard construction (Remark 1) and the q-power-mean construction.

Problem	Quantity	$q = 1$	$q = 2$	$q = 3$	$q = 4$	$q = 5$
4×3 (6 orderings)	Max. weight spread, standard construction	0.501	0.789	0.821	0.869	0.981
	Max. weight spread, q-power mean	0	0	0	0	0
	Distinct integrated rankings, standard	2	3	3	2	2
	Distinct integrated rankings, q-power mean	1	1	1	1	1
5×5 (120 orderings)	Max. weight spread, standard construction	0.232	0.564	0.633	0.717	0.787
	Max. weight spread, q-power mean	0	0	0	0	0
	Distinct integrated rankings, standard	19	26	26	26	26
	Distinct integrated rankings, q-power mean	1	1	1	1	1

6.2 Elicitation and agreement between models

All panel models returned five valid sessions except the Llama model in the enlarged problem (three in ten calls); every rejection was a q-rung violation in the performance judgments (Table 3). The between-model share of the score variance, averaged over judgments, is 0.680 for the preference and 0.723 for the performance judgments in the enlarged problem, and 0.558 and 0.476 in the reduced problem. The concordance of the model-wise rankings is $W = 0.752$ (enlarged) and 0.904 (reduced), whereas the rank correlations of the model-wise weights range down to -0.30 and -0.50. The panel weights are 0.1911, 0.2182, 0.2014, 0.2252 and 0.1641 (energy density, power density, cycle life, safety, technology readiness) and 0.3628, 0.3470 and 0.2901 (Performance, Safety, Durability). At $\tau = 0.1$ the consensus stage adjusted one (enlarged) and two (reduced) model-level relations by one fusion step; per-model results are given in Section S2.

6.3 What the LLM panel adds to the decision

Direct-data baseline

Replacing the LLM-elicited performance matrix by the decision table itself does not change the decision. Under every one of the 18 direct-data configurations, three mappings of the table combined with uniform, LLM-derived or knowledge-only weights under the two ranking variants, the integrated ranking is identical to that of the full LLM pipeline in both problems (Kendall's $\tau = 1$; Table S10): Li-ion $\succ$ Na-ion $\succ$ Zn-ion $\succ$ K-ion $\succ$ Ni-aq. in the enlarged and Zn-ion $\succ$ Li-ion $\succ$ Na-ion $\succ$ K-ion in the reduced problem. With the table available, the LLM stage therefore adds no decision-relevant information on these problems.

Knowledge-only condition

With the table in the prompt, the panel scores of the performance judgments reproduce the normalized table values (Pearson $r = 0.999$ and 0.998; Table 5, Fig. S8). Without the table they diverge from it ($r = 0.747$ and 0.886). In the enlarged problem the knowledge-only panel agrees with the table on energy density (Spearman $\rho = 1.00$), weakly on power density ($\rho = 0.50$), and inverts the order on cycle life ($\rho = -0.22$), judging the Li-ion capacitor best where the full-cell data place the Zn-ion device first. The integrated rankings change to Zn-ion $\succ$ Li-ion $\succ$ Ni-aq. $\succ$ Na-ion $\succ$ K-ion (enlarged, standard variant; Kendall's $\tau = 0.20$ with the main ranking) and Zn-ion $\succ$ Na-ion $\succ$ K-ion $\succ$ Li-ion (reduced; $\tau = 0.33$), and the concordance of the model-wise rankings falls to $W = 0.496$ and 0.408. The change is caused by the performance judgments: combining the knowledge-only weights with the table leaves the ranking unchanged. The criterion weights are stable across the two conditions, with a largest change of 0.016 (enlarged) and 0.032 (reduced); this stability does not establish that the weights are appropriate, since no human reference is available.

Table 5 Elicitation with and without the decision table in the prompt: agreement of the panel scores of the performance judgments with the normalized table values, concordance of the model-wise rankings, between-model shares, sessions and cost (six models including the anchor).

	Reduced (4 × 3)		Enlarged (5 × 5)	
Quantity	With table	Knowledge only	With table	Knowledge only
Pearson r, panel scores vs table	0.998	0.886	0.999	0.747
Spearman ρ, panel scores vs table	0.991	0.864	0.992	0.732
Kendall's W of model rankings	0.904	0.408	0.752	0.496
Between-model share β, H	0.558	0.745	0.680	0.537
Between-model share β, P	0.476	0.694	0.723	0.741
Accepted sessions / calls (six models)	30/31	30/30	26/40	30/32
Cost (USD)	1.148	1.191	2.131	1.905

6.4 Representation sufficiency

On the panel data, replacing every hesitant set by its q-power-mean element yields the same integrated ranking as the elicited sets in both problems, as Proposition 4

predicts; the weights coincide to four decimals in the enlarged problem and differ by at most 0.006 in the reduced problem, where one model-level relation performs a fusion step under the hesitant but not under the collapsed representation, so that assumption (iii) is violated. Under the conditional perturbation, the two representations give the same integrated ranking in 0.998–1.000 (enlarged) and 0.963–0.997 (reduced) of the draws. Representations by the smallest or largest degree of each tuple change the results: they exchange the third and fourth positions of the enlarged problem and reproduce the ranking of the hesitant representation in only 0.301–0.477 of the perturbed draws (Section S5). The hesitant sets thus add no ranking information beyond their q-power means in this case, but protect the results against the choice of a single endpoint. When a single central judgment is elicited per entry instead of hesitant sets, all sessions are accepted, the integrated rankings are identical to those of the hesitant panel in both problems (Kendall's $\tau = 1$), the weights differ by at most 0.0103 (reduced) and 0.0015 (enlarged), and the panel scores of the performance judgments correlate with those of the hesitant panel at $r = 0.9998$ and 0.9997 (Table S7). Together with Proposition 4, eliciting hesitant sets thus adds no decision information over a single central judgment in this case.

6.5 Ranking uncertainty and dependence on the ranking variant

Hierarchical resampling

Table 6 and Fig. 2 summarize the rank-1 probabilities. When models and sessions are resampled and the LLM-elicited performance judgments are used, the Li-ion capacitor is ranked first in 0.748 of the draws of the enlarged problem under the standard and in 1.000 under the fixed-bound variant, with the Na-ion capacitor as the only frequent competitor under the standard variant (0.167); the deterministic integrated ranking recurs in 0.290 and 0.556 of the draws. In the reduced problem, the Zn-ion device is first in 0.589 (standard) and 1.000 (fixed-bound) of the draws. When, in addition, the literature rows are resampled and the rebuilt table is used directly, the rankings disperse: the Li-ion capacitor is first in 0.533 and 0.504 of the draws, the Na-ion capacitor in 0.285 and 0.235, and the deterministic ranking recurs in 0.059 and 0.112. The weight standard deviations (0.0047–0.0181 and 0.0156–0.0280) exceed those of the conditional perturbation by factors of about 4 to 14. The source data contribute more ranking uncertainty than the panel, and, since single-source cells are not varied, this uncertainty is understated.

Method dependence

On the unperturbed data the two variants yield the same integrated ranking in both problems, but under every source of variation the probability of the top rank depends on the variant (Table 6). Under the conditional perturbation, the Li-ion and Na-ion capacitors are ranked first in 0.474 and 0.436 of the draws under the standard variant and the Li-ion capacitor in every draw under the fixed-bound variant; the apparent tie under the standard variant is therefore a property of the variant and of the perturbation model, not of the data alone. The variants also respond differently

Table 6 Probability of rank 1 and of the deterministic integrated ranking under each source of variation, for the standard (Std.) and fixed-bound (F.-b.) variants. P: performance matrix elicited from the LLMs or taken from the (rebuilt) decision table. Hierarchical: models and sessions, and in the second row also source rows, resampled jointly ($B = 2000$). The last row is the conditional perturbation of every elicited degree ($N = 10\,000$).

		4 × 3: Zn-ion		5 × 5: Li-ion		5 × 5: Na-ion	
Source of variation	P	Std.	F.-b.	Std.	F.-b.	Std.	F.-b.
Models, sessions (hierarchical)	LLM	0.589	1.000	0.748	1.000	0.167	0.000
Models, sessions, source rows (hierarchical)	table	0.560	0.716	0.533	0.504	0.285	0.235
Source rows only (bootstrap)	table	0.548	0.734	0.563	0.499	0.282	0.220
Degrees, uniform ±10 % (conditional)	LLM	0.718	0.993	0.474	1.000	0.436	0.000
Probability of the deterministic integrated ranking							
Models, sessions (hierarchical)	LLM	0.556	1.000	0.290	0.556		
Models, sessions, source rows (hierarchical)	table	0.118	0.126	0.059	0.112		
Source rows only (bootstrap)	table	0.097	0.141	0.055	0.146		

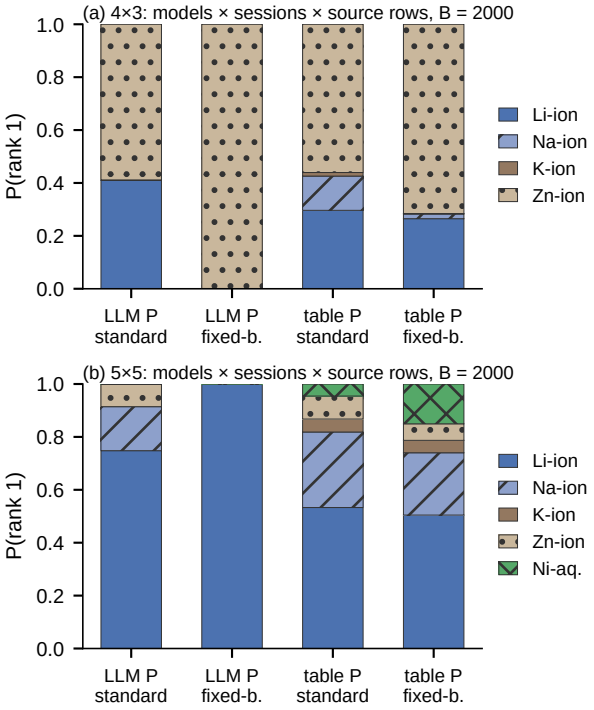


Fig. 2 Probability of rank 1 under hierarchical resampling ($B = 2000$) of models, sessions and source rows, with LLM-elicited (LLM P) or directly mapped (table P) performance values, for the standard and fixed-bound variants: (a) reduced, (b) enlarged problem.

to the encoding of the ordinal criteria: with the table used directly, none of the 16 encoding combinations of the enlarged problem changes the integrated ranking under the standard variant, whereas 14 change it under the fixed-bound variant, which uses the absolute normalized values (Table S20).

Full multiplicative form

The min–max rescaling (15) of the standard variant makes its full multiplicative component depend on ε: with the LLM-elicited performance matrix, its ranking changes between $\varepsilon \leq 10^{-6}$ and $\varepsilon \geq 10^{-3}$ in both problems, and with the table in the enlarged problem already between 10^{-9} and 10^{-6} (Table S24). The integrated ranking is unchanged for every ε, because the ratio system and the reference point outvote this component. With the transformation $(S + 1)/2$ of the fixed-bound variant no factor vanishes, and neither the component nor the integrated ranking depends on ε over 10^{-12} to 10^{-2}.

Data layer and rank reversal

Leaving out one source per cell changes the enlarged-problem ranking in 7 (standard) and 4 (fixed-bound) of 123 evaluable cases, and the top rank in 2 cases under each variant (removal of either of two Li-ion cycle-life sources); 11 single-source cases are not evaluable. Standard MULTIMOORA reverses ranks frequently: under perturbation of $\pm 10\,\%$ the removal test fails for the integrated ranking in 1.000 of the draws of the enlarged problem, whereas the fixed-bound variant fails in 2 of 80 000 draws over all settings (Section S7). Five hesitant and q-rung baselines agree on the first position except VIKOR, whose acceptable-advantage condition fails (Table S12).

6.6 Prompt-order effects on the elicited judgments

The elicited judgments depend on the order of the prompt. The between-order share of the judgment-score variance exceeds its permutation null in all four tests: 0.494 against a null mean of 0.281 (preference judgments, enlarged; $p = 0.0005$, the smallest attainable value), 0.364 against 0.243 (reduced; $p = 0.009$), and 0.375 against 0.282 and 0.408 against 0.243 for the performance judgments ($p = 0.0005$). In the reduced problem the Llama model has no accepted session in one order and is excluded from this comparison (Table S5). The top choice is unchanged under every order (Li-ion in the enlarged, Zn-ion in the reduced problem), but the lower positions of the enlarged problem shift (Kendall's $\tau = 0.6$, 0.8 and 1.0 against the original order), and the panel weights move by at most 0.034 (enlarged) and 0.049 (reduced; Table S6). The order invariance of Section 6.1 therefore holds for the aggregation but not for the pipeline as a whole.

7 Discussion

7.1 Aggregation artifacts and representation

The order dependence of the standard construction is a property of the operator and the reciprocity rebuild, not of the data; the q-power mean with row-sum weights removes it and the zero weight of the min-shift crispification. This guarantee is limited to the aggregation of judgments already elicited: the prompt-order experiment shows order effects in the judgments themselves, so that criterion and alternative orders should be randomized or balanced across sessions, as aggregation over a full factorial design of orders has been reported to remove response-order bias in other judgment

tasks [41]. Proposition 4 and the representation ablation show that, on these data, the dual hesitant sets carry no ranking information beyond their q-power means. A directly elicited single central judgment gives the same rankings, so that representation sufficiency is a finding in its own right: in this case, hesitant elicitation adds no decision information and is optional, whereas reducing elicited sets to one of their endpoints changes the results.

7.2 What LLM panels add

The direct-data baseline and the knowledge-only condition delimit the contribution of the panel. When the decision table is available, the LLM-elicited performance judgments reproduce it and the decision is the one obtained from the table directly, with uniform as well as with LLM-derived weights; on these problems, the panel then adds cost and variance but no decision-relevant information. When the table is withheld, the panel changes the decision through performance judgments that diverge from the full-cell data on cycle life and power density; as an untested interpretation, the preference for the Li-ion capacitor on cycle life may reflect the maturity of this class, the only one with an in-vehicle demonstration in the extracted data. The criterion-importance judgments are stable across both conditions and across resampling, but stability is not validity: whether these weights represent the priorities of vehicle engineers cannot be determined without a human panel. In a renewable-energy policy problem, LLM-generated criterion weights agreed closely with those of domain experts, whereas the decision matrices and rankings differed more [4]; whether this pattern holds for the present problem is not known. The defensible role of the panel is therefore that of a documented source of criterion-importance judgments, not a substitute for performance data.

7.3 Uncertainty and the choice of ranking variant

The uncertainty statements rest on the hierarchical resampling: with LLM-elicited performance judgments it favors the Li-ion capacitor in the enlarged problem under both variants, and with resampled source data the first position is shared, with the Li-ion capacitor most probable. The apparent tie under the conditional perturbation depends on the assumed levels and the variant, and the choice of variant changes the rank-1 probabilities even where the deterministic rankings coincide. The fixed-bound variant avoids rank reversal and the ε dependence but responds strongly to the spacing of the ordinal scales; the standard variant is insensitive to the encoding but set dependent. Neither variant is preferred on these grounds alone, and no technology statement is made beyond the probabilities reported under each.

7.4 Limitations

The data layer rests on few sources per cell; three alternatives have single-source energy and power values that the row bootstrap does not vary, the cycle counts come from tests with differing end points, cost is absent, and the technology readiness level is a single assignment per alternative. The LLMs may have been trained on the same

literature from which the table was built, so that the knowledge-only condition separates the judgments from the supplied data but not from the published sources. The main elicitation used a single fixed order of criteria and alternatives, which the prompt-order experiment shows to influence the judgments; two panel models were sampled at the provider default temperature; and model versions behind a fixed identifier can change. No human-expert panel was elicited; the LLM-elicited weights and rankings are not validated, and the agreement reported is agreement among models. An elicitation of domain experts with the identical protocol is the principal item of future work.

8 Conclusion

Three results of this study are methodological. The standard construction of the group relation depends on the order of the criteria and produced up to 26 distinct rankings on the same judgments, whereas the q-power mean with row-sum priority weights is provably order invariant at the aggregation level and keeps every weight positive; the elicited judgments themselves, however, shift with the order of the prompt, so that orders should be randomized across sessions. The results depend on each hesitant set only through the means of the q-th powers of its degrees in the absence of consensus adjustments, projections and padding, and a directly elicited single central judgment yields the same rankings, so that hesitant elicitation is optional in the case examined. The rank-1 probabilities depend on the choice between standard and fixed-bound MULTIMOORA, and the min–max step of the standard full multiplicative form introduces a dependence on ε that the transformation $(S + 1)/2$ removes.

On the contribution of the panel, the direct-data baseline showed that, with the decision table available, the LLM-elicited performance judgments reproduce the decision obtained from the table directly; without the table, the panel changes the decision through performance judgments that diverge from the published evidence. The criterion-importance judgments were stable, but their usefulness remains unvalidated. Under hierarchical resampling, the Li-ion capacitor leads the enlarged problem when the LLM-elicited performance judgments are used, but the first position is shared once the literature source rows are resampled; the source data contribute more ranking uncertainty than the panel. The principal extensions are a human-expert panel elicited with the identical protocol, a data layer with more sources per cell and cost data, and repeated elicitations over successive model versions.

Supplementary Information. The online version contains supplementary material (Online Resource 1, file `ESM_1.pdf`): the elicitation prompts, the per-model results, the prompt-order and single-judgment controls, the proofs of Propositions 1–4 and the weighting algorithm, the representation ablation, the direct-data and hesitant baselines, the robustness analyses, the data-layer analyses and additional figures.

Acknowledgements. Communication of this research is made possible through monetary assistance by Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

Statements and Declarations

Funding. The communication of this research was supported by Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office through Publication Fund E15216.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable. The study involves no human participants, human data or animals; the elicited judgments were produced by large language models accessed through their programming interfaces.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The elicitation prompts, the raw responses of every model call (accepted and rejected), the extracted literature data with source identifiers and the decision tables are deposited at Zenodo under the DOI https://doi.org/10.5281/zenodo.23695939 (released under CC BY 4.0).

Code availability. The code of the framework and of all experiments, and the scripts that generate every table and figure of this article, are included in the same Zenodo deposit (released under the MIT licence).

Materials availability. Not applicable.

Author contributions. [To be completed by the authors: contribution of each author.] All authors read and approved the final manuscript.

References

[1] Lu, X., Li, J., Takeuchi, K., Kashima, H.: AHP-powered LLM reasoning for multi-criteria evaluation of open-ended responses. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 1847–1856. Association for Computational Linguistics, Miami, FL, USA (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.101

[2] Park, H., Oh, H., Gao, F., Kwon, O.: Enhancing analytic hierarchy process modelling under uncertainty with fine-tuning LLM. Expert Syst. **42**(6), 70051 (2025) https://doi.org/10.1111/exsy.70051

[3] Vahidnia, M.H.: Multi-agent systems of large language models as weight assigners: An approach to collaborative weighting in spatial multi-criteria decision-making. Geomatica **77**(2), 100071 (2025) https://doi.org/10.1016/j.geomat.2025.100071

[4] Więckowski, J., Paradowski, B., Kizielewicz, B., Bouraima, M.B.: Evaluating renewable energy development policies: A SWOT-based multi-criteria decision

analysis comparison of domain experts and large language models. Appl. Soft Comput. **204**, 116417 (2027) https://doi.org/10.1016/j.asoc.2026.116417

[5] Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., Sui, Z.: Large language models are not fair evaluators. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9440–9450. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.acl-long.511

[6] Yin, H., Vardi, S., Choudhary, V.: Fragile preferences: A deep dive into order effects in large language models. PNAS Nexus **5**(8), 246 (2026) https://doi.org/10.1093/pnasnexus/pgag246

[7] Yager, R.R.: Generalized orthopair fuzzy sets. IEEE Trans. Fuzzy Syst. **25**(5), 1222–1230 (2017) https://doi.org/10.1109/TFUZZ.2016.2604005

[8] Xu, Y., Shang, X., Wang, J., Wu, W., Huang, H.: Some q-rung dual hesitant fuzzy Heronian mean operators with their application to multiple attribute group decision-making. Symmetry **10**(10), 472 (2018) https://doi.org/10.3390/sym10100472

[9] Xu, Y., Liu, S., Wang, J., Shang, X.: Consensus checking and improving methods for AHP with q-rung dual hesitant fuzzy preference relations. Expert Syst. Appl. **208**, 117902 (2022) https://doi.org/10.1016/j.eswa.2022.117902

[10] Brauers, W.K.M., Zavadskas, E.K.: Project management by MULTIMOORA as an instrument for transition economies. Technol. Econ. Dev. Econ. **16**(1), 5–24 (2010) https://doi.org/10.3846/tede.2010.01

[11] Wang, X., Wei, J., Schuurmans, D., Le, Q.V., Chi, E.H., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations (ICLR 2023). OpenReview.net, Kigali, Rwanda (2023). https://openreview.net/forum?id=1PL1NIMMrw

[12] Park, P.S., Schoenegger, P., Zhu, C.: Diminished diversity-of-thought in a standard large language model. Behav. Res. Methods **56**(6), 5754–5770 (2024) https://doi.org/10.3758/s13428-023-02307-x

[13] Schoenegger, P., Tuminauskaite, I., Park, P.S., Bastos, R.V.S., Tetlock, P.E.: Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. Sci. Adv. **10**(45), 1528 (2024) https://doi.org/10.1126/sciadv.adp1528

[14] Liu, P., Wang, P.: Some q-rung orthopair fuzzy aggregation operators and their applications to multiple-attribute decision making. Int. J. Intell. Syst. **33**(2), 259–280 (2018) https://doi.org/10.1002/int.21927

[15] Li, L., Ji, C., Wang, J.: A novel multi-attribute group decision-making method based on q-rung dual hesitant fuzzy information and extended power average operators. Cogn. Comput. **13**(5), 1345–1362 (2021) https://doi.org/10.1007/s12559-021-09932-8

[16] Kou, Y., Feng, X., Wang, J.: A novel q-rung dual hesitant fuzzy multi-attribute decision-making method based on entropy weights. Entropy **23**(10), 1322 (2021) https://doi.org/10.3390/e23101322

[17] Sarkar, A., Moslem, S., Esztergár-Kiss, D., Akram, M., Jin, L., Senapati, T.: A hybrid approach based on dual hesitant q-rung orthopair fuzzy Frank power partitioned Heronian mean aggregation operators for estimating sustainable urban transport solutions. Eng. Appl. Artif. Intell. **124**, 106505 (2023) https://doi.org/10.1016/j.engappai.2023.106505

[18] Li, T., Zhang, L.: Group decision-making based on multiplicative consistency and consensus of linguistic interval-valued q-rung orthopair fuzzy preference relations. Comput. Appl. Math. **44**(2), 53 (2025) https://doi.org/10.1007/s40314-024-03020-x

[19] Tanino, T.: Fuzzy preference orderings in group decision making. Fuzzy Sets Syst. **12**(2), 117–131 (1984) https://doi.org/10.1016/0165-0114(84)90032-0

[20] Fedrizzi, M., Brunelli, M.: On the priority vector associated with a reciprocal relation and a pairwise comparison matrix. Soft Comput. **14**(6), 639–645 (2010) https://doi.org/10.1007/s00500-009-0432-2

[21] Mohammed, R.T., Zaidan, A.A., Yaakob, R., Sharef, N.M., Abdullah, R.H., Zaidan, B.B., Albahri, O.S., Abdulkareem, K.H.: Determining importance of many-objective optimisation competitive algorithms evaluation criteria based on a novel fuzzy-weighted zero-inconsistency method. Int. J. Inf. Technol. Decis. Mak. **21**(1), 195–241 (2022) https://doi.org/10.1142/S0219622021500140

[22] Qahtan, S., Sharif, K.Y., Zaidan, A.A., Alsattar, H.A., Albahri, O.S., Zaidan, B.B., Zulzalil, H., Osman, M.H., Alamoodi, A.H., Mohammed, R.T.: Novel multi security and privacy benchmarking framework for blockchain-based IoT healthcare industry 4.0 systems. IEEE Trans. Ind. Inform. **18**(9), 6415–6423 (2022) https://doi.org/10.1109/TII.2022.3143619

[23] Alsalem, M.A., Alamoodi, A.H., Albahri, O.S., Albahri, A.S., Martínez, L., Yera, R., Duhaim, A.M., Sharaf, I.M.: Evaluation of trustworthy artificial intelligent healthcare applications using multi-criteria decision-making approach. Expert Syst. Appl. **246**, 123066 (2024) https://doi.org/10.1016/j.eswa.2023.123066

[24] Qahtan, S., Alsattar, H.A., Zaidan, A.A., Deveci, M., Pamucar, D., Delen, D.: Performance assessment of sustainable transportation in the shipping industry

using a q-rung orthopair fuzzy rough sets-based decision making methodology. Expert Syst. Appl. **223**, 119958 (2023) https://doi.org/10.1016/j.eswa.2023.119958

[25] Xiao, L., Fang, T., Huang, G., Deveci, M.: An integrated design concept evaluation method based on fuzzy weighted zero inconsistency and combined compromise solution considering inherent uncertainties. Adv. Eng. Inform. **65**, 103097 (2025) https://doi.org/10.1016/j.aei.2024.103097

[26] Ghailani, H., Zaidan, A.A., Qahtan, S., Alsattar, H.A., Al-Emran, M., Deveci, M., Delen, D.: Developing sustainable management strategies in construction and demolition wastes using a q-rung orthopair probabilistic hesitant fuzzy set-based decision modelling approach. Appl. Soft Comput. **145**, 110606 (2023) https://doi.org/10.1016/j.asoc.2023.110606

[27] Hafezalkotob, A., Hafezalkotob, A., Liao, H., Herrera, F.: An overview of MULTIMOORA for multi-criteria decision-making: Theory, developments, applications, and challenges. Inf. Fusion **51**, 145–177 (2019) https://doi.org/10.1016/j.inffus.2018.12.002

[28] Zhang, C., Chen, C., Štreimikienė, D., Baležentis, T.: Intuitionistic fuzzy MULTIMOORA approach for multi-criteria assessment of the energy storage technologies. Appl. Soft Comput. **79**, 410–423 (2019) https://doi.org/10.1016/j.asoc.2019.04.008

[29] Rani, P., Mishra, A.R.: Fermatean fuzzy Einstein aggregation operators-based MULTIMOORA method for electric vehicle charging station selection. Expert Syst. Appl. **182**, 115267 (2021) https://doi.org/10.1016/j.eswa.2021.115267

[30] Mishra, A.R., Rani, P., Pamučar, D., Hezam, I.M., Saha, A.: Entropy and discrimination measures based q-rung orthopair fuzzy MULTIMOORA framework for selecting solid waste disposal method. Environ. Sci. Pollut. Res. **30**(5), 12988–13011 (2023) https://doi.org/10.1007/s11356-022-22734-1

[31] Gao, F.: An integrated multi criteria decision making method using dual hesitant fuzzy sets with application for unmanned aerial vehicle selection. Sci. Rep. **15**(1), 12637 (2025) https://doi.org/10.1038/s41598-025-95981-0

[32] Alhadithi, O.S., Audah, L., Alhartomi, M.A., Hamdi, M.M., Hirsi, A., Ansa, G.O., Farah, A.: Enhanced multi-criteria decision-making framework for hybrid super-capacitor selection in electric vehicles using FWZIC and fuzzy MULTIMOORA. Int. J. Intell. Eng. Syst. **19**(11), 986–1009 (2026) https://doi.org/10.22266/ijies2026.1130.51

[33] Saluja, R.S., Singh, V.: An improved fuzzy MULTIMOORA approach and its application in welding process selection. Int. J. Fuzzy Syst. **25**(4), 1707–1726 (2023) https://doi.org/10.1007/s40815-023-01472-7

[34] Farias Aires, R.F., Ferreira, L.: The rank reversal problem in multi-criteria decision making: A literature review. Pesqui. Oper. **38**(2), 331–362 (2018) https://doi.org/10.1590/0101-7438.2018.038.02.0331

[35] Wang, X., Triantaphyllou, E.: Ranking irregularities when evaluating alternatives by using some ELECTRE methods. Omega **36**(1), 45–63 (2008) https://doi.org/10.1016/j.omega.2005.12.003

[36] Jiang, L., Liao, H., De Baets, B.: A simulation study on the probabilities of rank reversal, tie making, and tie breaking for multiple criteria decision making methods. Omega **125**, 103033 (2024) https://doi.org/10.1016/j.omega.2023.103033

[37] Trillo, J.R., Martínez, M.Á., Zadrożny, S., Kacprzyk, J., Herrera-Viedma, E., Cabrerizo, F.J.: A group decision-making approach leveraging preference relations derived from large language model. Procedia Comput. Sci. **242**, 452–459 (2024) https://doi.org/10.1016/j.procs.2024.08.161

[38] González-Quesada, J.C., Trillo, J.R., Porcel, C., Pérez, I.J., Cabrerizo, F.J.: Modelling large-scale group decision-making through grouping with large language models. Future Internet **17**(9), 381 (2025) https://doi.org/10.3390/fi17090381

[39] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track, pp. 46595–46623. Curran Associates, Inc., Red Hook, NY, USA (2023). https://doi.org/10.52202/075280-2020

[40] Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., Vosoughi, S.: Judging the judges: A systematic study of position bias in LLM-as-a-judge. In: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pp. 292–314. The Asian Federation of Natural Language Processing and the Association for Computational Linguistics, Mumbai, India (2025). https://doi.org/10.18653/v1/2025.ijcnlp-long.18

[41] Brucks, M., Toubia, O.: Prompt architecture induces methodological artifacts in large language models. PLoS One **20**(4), 0319159 (2025) https://doi.org/10.1371/journal.pone.0319159

[42] Haldar, R., Hockenmaier, J.: Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In: Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 24986–25004. Association for Computational Linguistics, Suzhou, China (2025). https://doi.org/10.18653/v1/2025.findings-emnlp.1361

[43] Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., Hooi, B.: Can

LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In: The Twelfth International Conference on Learning Representations (ICLR 2024). OpenReview.net, Vienna, Austria (2024). https://openreview.net/forum?id=gjeQKFxFpZ

[44] Savage, T., Wang, J., Gallo, R., Boukil, A., Patel, V., Safavi-Naini, S.A.A., Soroush, A., Chen, J.H.: Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. J. Am. Med. Inform. Assoc. **32**(1), 139–149 (2025) https://doi.org/10.1093/jamia/ocae254

[45] Zhong, C., Deng, Y., Hu, W., Qiao, J., Zhang, L., Zhang, J.: A review of electrolyte materials and compositions for electrochemical supercapacitors. Chem. Soc. Rev. **44**(21), 7484–7539 (2015) https://doi.org/10.1039/C5CS00303B

[46] Gond, R., van Ekeren, W., Mogensen, R., Naylor, A.J., Younesi, R.: Non-flammable liquid electrolytes for safe batteries. Mater. Horiz. **8**(11), 2913–2928 (2021) https://doi.org/10.1039/D1MH00748C

[47] Gong, X., Chen, J., Lee, P.S.: Zinc-ion hybrid supercapacitors: Progress and future perspective. Batter. Supercaps **4**(10), 1529–1546 (2021) https://doi.org/10.1002/batt.202100034

[48] Mankins, J.C.: Technology readiness assessments: A retrospective. Acta Astronaut. **65**(9–10), 1216–1223 (2009) https://doi.org/10.1016/j.actaastro.2009.03.058

[49] Bruno, I., Lobo, G., Covino, B.V., Donarelli, A., Marchetti, V., Schiavone Panni, A., Molinari, F.: Technology readiness revisited: a proposal for extending the scope of impact assessment of European public services. In: Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2020), pp. 369–380. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3428502.3428552

[50] Liu, X., Sun, Y., Tong, Y., Wang, X., Zheng, J., Wu, Y., Li, H., Niu, L., Hou, Y.: Exploration in materials, electrolytes and performance towards metal ion (Li, Na, K, Zn and Mg)-based hybrid capacitors: A review. Nano Energy **86**, 106070 (2021) https://doi.org/10.1016/j.nanoen.2021.106070

[51] Lamb, J.J., Burheim, O.S.: Lithium-ion capacitors: A review of design and active materials. Energies **14**(4), 979 (2021) https://doi.org/10.3390/en14040979

[52] Tang, H., Yao, J., Zhu, Y.: Recent developments and future prospects for zinc-ion hybrid capacitors: a review. Adv. Energy Mater. **11**(14), 2003994 (2021) https://doi.org/10.1002/aenm.202003994