University of New Mexico

Article

Typed Abstention for LLM-Elicited Fuzzy Criteria in Multicriteria Transportation Decisions

Maikel Leyva-Vázquez 1,2,3,* and Nezir Aydin 4

1 Universidad Bernardo O'Higgins, Santiago, Chile

2 Universidad Bolivariana del Ecuador, Durán, Ecuador; myleyvav@ube.edu.ec

3 Universidad de Guayaquil, Guayaquil, Ecuador

4 College of Science and Engineering, Hamad Bin Khalifa University, Qatar; naydin@hbku.edu.qa

* Correspondence: mleyvaz@gmail.com

Abstract: Fuzzy multicriteria decision-making (MCDM) increasingly draws its criterion evaluations from large language models (LLMs), but the aggregation step typically treats these elicited judgments as interchangeable with expert opinion, with no mechanism to flag disagreement, insufficiency, or unreliable sources. We port a source-preserving, typed-abstention decision policy into fuzzy-MCDM pipelines where criteria are elicited from LLMs, instantiated on a four-alternative sustainable-transportation case, and report two live pilots. Pilot 1 (96 calls; three well-aligned LLMs; criteria with more directly specified operational anchors) returns all-coherent: cross-model dispersion (mean 0.56 of 4 ordinal levels) never crosses the pre-specified conflict/opposition thresholds, so the gate withholds nothing and gated and undifferentiated TOPSIS rankings coincide. We then redesign the case — criteria with greater normative and contextual dependence and a panel that adds a smaller open model — and rerun it as Pilot 2 (128 calls). The gate now fires on real disagreement: 3 of 16 cells are typed conflicting, 2 of 4 criteria are dropped from the gated ranking, and the gated and naive TOPSIS scores diverge materially even where the ranking order coincides. We report both pilots candidly, verify the gate's logic with synthetic unit tests, and specify what a domain-expert reference ranking would still need to add before the framework's coverage/error trade-off can be evaluated against ground truth.

Keywords: fuzzy multicriteria decision-making; typed abstention; large language models as elicitation sources; TOPSIS; source-preserving aggregation; sustainable transportation; evidence theory

1. Introduction

Multicriteria decision-making (MCDM) methods — TOPSIS, VIKOR, AHP, and their fuzzy and hesitant-fuzzy extensions — are the standard tool for ranking alternatives against several, often conflicting, criteria [1]. A recent review identifies artificial intelligence, and specifically AI-generated criterion evaluations, as a rising input source expected to reduce the cost of expert elicitation in fuzzy MCDM pipelines for sustainable transportation planning [2]. That review, like the wider literature it

surveys, treats an AI-generated evaluation as functionally equivalent to a human expert's linguistic judgment once it has been fuzzified: both feed the same aggregation step, and neither carries a flag for cases in which the evaluation is unreliable, internally contradictory, or should not have been elicited from that source at all.

This is a consequential gap once the AI source is a large language model (LLM), known to produce fluent evaluations even when the underlying evidence is absent or contradictory [3]. We address this gap by porting a source-preserving decision policy, originally developed for panel aggregation under conflicting or insufficient evidence [4], into the fuzzy MCDM setting where LLMs act as elicitation sources. We report two real, live pilots rather than a purely conceptual proposal. The first came back all-coherent — an honest null result we report rather than discard. We then deliberately redesigned the case to test whether the gate can fire on real data at all: contested normative criteria plus a weaker model in the panel. It does fire, in a way that changes the ranking's evidentiary basis even where it does not change the ranking's order.

2. Materials and Methods

2.1. Related Work

Fuzzy extensions of TOPSIS [11], VIKOR, and AHP are well established for ranking alternatives under linguistic, imprecise criterion evaluations, with recent applications in sustainable transportation planning [2], hesitant-fuzzy evaluation of telecommunication operators [1], and multi-objective facility-location design for pandemic healthcare logistics [5]. These studies elicit criterion weights and alternative scores, fuzzify them, and aggregate by a fixed rule, validated by sensitivity analysis rather than by any check on whether the input panel itself was fit to aggregate. Separately, Dempster-Shafer theory and its extensions [12] formalize the combination of independent, possibly conflicting evidence sources — including, within MCDM itself, evidential-reasoning variants of outranking methods such as DS-ELECTRE [13] — and hallucination-detection work shows that a single softmax-normalized entailment score structurally cannot represent paraconsistent evidence [3]. Typed abstention generalizes this to a small taxonomy of reasons for withholding a decision. Prior work on LLM evaluation bias shows that judgments are sensitive to factors irrelevant to the decision — institutional prestige cues shift model outputs by an amount comparable to, and in some domains larger than, geographic cues [6] — direct evidence that LLM-elicited MCDM criteria can carry structured, source-dependent bias that naive aggregation propagates silently. A narrative review of neutrosophy in medical sciences documents how explicit indeterminacy has been used across clinical decision and medical-imaging problems [18].

Separately from MCDM, the LLM-as-judge and uncertainty-quantification literature has begun to formalize exactly the two disagreement signals we use: within-model self-consistency across repeated queries and cross-model disagreement between independent judges. Hamidieh et al. [8] show that cross-model semantic disagreement is a distinct, complementary uncertainty term to self-consistency, most informative precisely when self-consistency is uninformative; Liu [9] shows that cross-model consensus, taken as agreement structure rather than as a score one model assigns another, selects correct answers better than self-consistency alone; Badshah et al. [10] calibrate an abstention threshold for pairwise LLM judging with a formal risk guarantee. Our typing rule (§2.4) is best read as porting this general self-consistency/cross-model-disagreement distinction into fuzzy-

MCDM criterion elicitation specifically — the novelty we claim is the application and the source-preserving MCDM policy it plugs into [4], not the underlying distinction between the two disagreement types, which these three papers establish independently and roughly concurrently with this work. The source-preserving gate also has a causal-knowledge analogue: medical-informatics work represents causal knowledge with directed graphs and fuzzy cognitive maps instead of collapsing interacting evidence into independent scores [17].

2.2. Source-Preserving Aggregation and the Typed-Abstention Gate

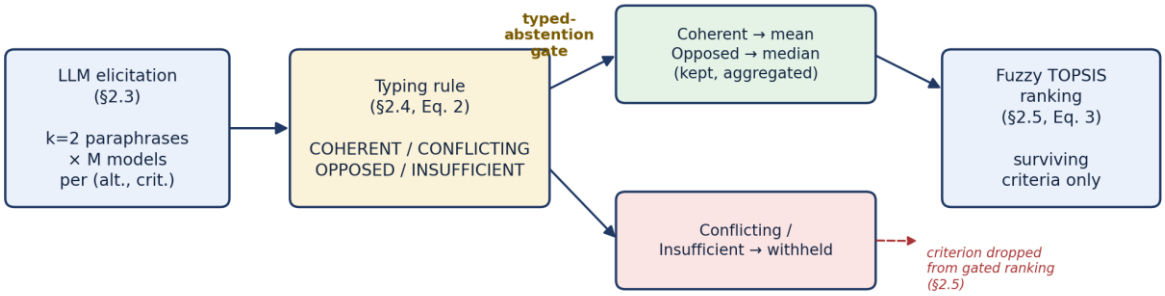

Figure 1. The typed-abstention pipeline: LLM elicitation feeds the typing rule (Eq. 2), which routes each cell to aggregation (coherent/opposed) or withholding (conflicting/insufficient); only surviving criteria reach the TOPSIS ranking (Eq. 3).

We adopt the policy of [4]: criterion aggregation happens within each source, and aggregation across sources is postponed until a decision to rank is made. Each source is classified before ranking as coherent, internally conflicting, opposed to other sources, or insufficient. Fuzzy criterion evaluations are elicited as one of five linguistic terms t ∈ {VL, L, M, H, VH}, each a triangular fuzzy number (a_t, b_t, c_t); the crisp, defuzzified value used downstream is its centroid:

$$centroid(t) = \frac{a_t + b_t + c_t}{3} \quad (1)$$

Table 1. Triangular fuzzy numbers and centroids for the five linguistic terms (Eq. 1).

t	VL	L	M	H	VH
(a_t, b_t, c_t)	(0, 0, .25)	(0, .25, .5)	(.25, .5, .75)	(.5, .75, 1)	(.75, 1, 1)
centroid(t)	0.083	0.25	0.5	0.75	0.917

2.3. LLM Elicitation Protocol

For alternatives A and criteria C, each LLM is queried k = 2 times per (alternative, criterion) cell under two paraphrased prompt templates, requesting a fuzzy linguistic term with a one-sentence justification, returned as JSON. Repeated, paraphrased elicitation from the same model exposes within-model (internal-conflict) disagreement; distinct model families expose cross-model

(opposition) disagreement. Per model, the representative level for a cell is the mean of its parseable ordinal levels (VL=0 … VH=4) — i.e., levels are averaged first and the centroid of Eq. 1 is applied afterward by linear interpolation between the two nearest terms' centroids, not the other way around; averaging the five centroid values directly would give a slightly different number. In both pilots both calls were parseable for every model, so the representative level is always the mean of exactly two ordinal levels; if only one call had been parseable, that single level would have been used and internal-conflict could not have been evaluated for that model on that cell.

2.4. Typing Rule and Aggregation

Let $l_{m,1}$ and $l_{m,2}$ denote model m's two ordinal levels (VL = 0 … VH = 4) for a cell, and let $\bar{l}_m$ denote their mean. A model is internally conflicting if its two calls differ by at least the threshold τc, fixed before inspecting any data:

$$conflict_m = 1 \Leftrightarrow |l_{m,1} - l_{m,2}| \geq \tau_c,\ \tau_c = 2 \quad (2)$$

A cell is typed INSUFFICIENT if fewer than two models return a parseable term; CONFLICTING if any usable model is internally conflicting; OPPOSED if the spread between usable models' representative levels is at least τo = 3 levels; otherwise COHERENT. Coherent cells aggregate by the mean of usable models' representative levels; opposed cells resolve by the median (the declared coalition rule); conflicting and insufficient cells are withheld from the gated pipeline. An undifferentiated ("naive") baseline instead averages every parseable call regardless of typing.

2.5. Ranking

A criterion enters the gated TOPSIS ranking only if every alternative has a non-withheld gated value for it; a single conflicting or insufficient cell anywhere in that criterion's column removes the whole criterion from the gated ranking, not just the affected alternative — this all-or-nothing column rule, not a per-alternative penalty, is what separates coverage at the cell level from coverage at the criterion level in §3.2/§3.3. Surviving criteria (equal weights $w_j = 1/n$) are ranked by crisp TOPSIS. For alternative i, with vector-normalized, weighted distances to the ideal (D_i^+) and anti-ideal (D_i^-) solutions, the closeness coefficient is:

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \quad (3)$$

3. Results

3.1. Case Definition

Illustrative decision problem, extending the case family of [2]: ranking four transportation-investment alternatives for a mid-sized Latin American city (framed in the prompts as Guayaquil, Ecuador) — Bus Rapid Transit (BRT), Light Rail Transit (LRT), an E-bike/micromobility network, and a fixed-route minibus system.

3.2. Pilot 1: Technical Criteria (Environmental, Economic, Social, Operational)

Models: anthropic/claude-haiku-4.5, openai/gpt-4o-mini, google/gemini-3.5-flash-lite, via the OpenRouter API in a single session (temperature 0.7; call-level timestamps are not retained in the archived JSON, so only the script-run date is claimed, not a per-call timestamp). 4 alternatives × 4 criteria × 3 models × 2 paraphrase variants = 96 calls; all 96 returned a parseable term.

Of 16 cells: 16 COHERENT, 0 CONFLICTING, 0 OPPOSED, 0 INSUFFICIENT. Cross-model dispersion averaged 0.562 of a possible 4 (max 1.000) — real disagreement, but below both thresholds for every cell. Gated coverage was therefore 100% and all four criteria survived. Table 2 gives the aggregated evaluation matrix (identical for the naive and gated pipelines here); Table 3 the resulting TOPSIS scores.

Table 2. Pilot 1 aggregated fuzzy-criterion levels (0=VL … 4=VH), gate = naive here.

Alternative	ENV	ECO	SOC	OPS
Bus Rapid Transit (BRT)	3.17	3.50	3.33	3.17
Light Rail Transit (LRT)	3.00	1.83	2.67	2.67
E-bike / Micromobility Network	4.00	3.50	2.33	1.83
Fixed-Route Minibus System	2.67	2.67	2.67	2.00

Table 3. Pilot 1 TOPSIS scores (naive = gated).

Rank	Alternative	Naive score	Gated score
1	Bus Rapid Transit (BRT)	0.8220	0.8220
2	E-bike / Micromobility Network	0.5005	0.5005
3	Light Rail Transit (LRT)	0.3743	0.3743
4	Fixed-Route Minibus System	0.3376	0.3376

Because the gate withheld nothing, the two pipelines produce the identical ranking (BRT > E-bike network > LRT > minibus). This is the expected outcome when LLM sources converge below the disagreement thresholds, not a general claim that naive and gated pipelines always agree.

We separately unit-tested the typing rule (§2.4) on four synthetic, non-experimental inputs — coherent, internally conflicting, opposed, and insufficient — confirming it correctly types and withholds each case when it occurs. This is a code-correctness check, not an experimental finding.

3.3. Pilot 2: Contested Criteria and a Weaker Model

Pilot 1 could not show the gate acting on real data. We redesigned the case along two axes, decided in advance: (i) criteria replaced with four normatively contested ones — fairness to displaced informal-sector drivers (DISP), gentrification risk (GENT, higher = lower risk), appropriateness of reallocating road space from cars (CARSPACE), and political acceptability to a fiscally conservative council (CULTFIT); (ii) the panel extended with a smaller, weaker open model, meta-llama/llama-3.1-8b-instruct [7], alongside the three from Pilot 1. Everything else — alternatives, scale, k = 2, thresholds — is unchanged. 4 × 4 × 4 × 2 = 128 calls, all parseable, 0 refusals.

Of 16 cells: 13 COHERENT, 3 CONFLICTING, 0 OPPOSED, 0 INSUFFICIENT. All three flagged cells were driven by a model's own two paraphrased answers landing far enough apart to cross τ_c — mostly the smaller Llama model, once joined by Gemini 3.5 Flash-Lite (Table 4). Substantively, the

criteria this knocked out of the gated ranking (DISP, CULTFIT) are the two most politically loaded of the four — a real, computed pattern, not a designed-in outcome.

Table 4. Pilot 2 cells flagged by the typed-abstention gate.

Alternative	Criterion	Type	Per-model representative levels
Bus Rapid Transit (BRT)	CULTFIT	CONFLICTING	claude-haiku-4.5=3.0; gpt-4o-mini=3.0; gemini-3.5-flash-lite=3.0; llama-3.1-8b-instruct=2.0
Light Rail Transit (LRT)	CULTFIT	CONFLICTING	claude-haiku-4.5=1.0; gpt-4o-mini=2.0; gemini-3.5-flash-lite=1.0; llama-3.1-8b-instruct=2.0
Fixed-Route Minibus System	DISP	CONFLICTING	claude-haiku-4.5=1.0; gpt-4o-mini=1.0; gemini-3.5-flash-lite=1.5; llama-3.1-8b-instruct=2.0

Table 5. Pilot 2 naive (gate = OFF) evaluation matrix.

Alternative	DISP	GENT	CARSPACE	CULTFIT
Bus Rapid Transit (BRT)	1.38	1.62	3.00	2.75
Light Rail Transit (LRT)	1.00	1.50	2.75	1.50
E-bike / Micromobility Network	1.00	1.62	2.75	2.88
Fixed-Route Minibus System	1.38	1.75	2.62	2.38

Table 6. Pilot 2 gated evaluation matrix (— = withheld).

Alternative	DISP	GENT	CARSPACE	CULTFIT
Bus Rapid Transit (BRT)	1.38	1.62	3.00	—
Light Rail Transit (LRT)	1.00	1.50	2.75	—
E-bike / Micromobility Network	1.00	1.62	2.75	2.88
Fixed-Route Minibus System	—	1.75	2.62	2.38

Table 7. Pilot 2 TOPSIS scores, naive vs. gated (GENT + CARSPACE only).

Rank	Alternative	Naive score	Gated score
1	Bus Rapid Transit (BRT)	0.8704	0.6687
2	Fixed-Route Minibus System	0.6706	0.5329
3	E-bike / Micromobility Network	0.6315	0.4295
4	Light Rail Transit (LRT)	0.0629	0.2015

The ranking order coincides (BRT > minibus > e-bike network > LRT), but the scores do not: BRT's naive score (0.870) drops to its gated score (0.669), and LRT's naive score (0.063) rises to its gated score (0.201). Per the all-or-nothing column rule (§2.5), this is not attributable to one criterion's judgments in isolation: CULTFIT is dropped because of conflicting cells in both BRT and LRT, DISP is dropped because of a conflicting cell in the minibus alternative, and TOPSIS is then fully recomputed over only the two surviving criteria (GENT, CARSPACE) with renormalized distances — a decomposition we verified by recomputing TOPSIS on each three-criterion subset separately: dropping CULTFIT alone moves BRT by -0.051 and dropping DISP alone by -0.017, neither of which explains the full -0.202 gated shift, confirming the change comes from the joint column removal and renormalization rather than from excluding one criterion's noisy judgments. This is the coverage/reliability trade-off the framework promises, observed on real model output: the gate changes what evidence and what criterion set a decision-maker sees, even where it does not flip the final order.

4. Discussion

Read together, the two pilots say something a single one could not. Pilot 1 shows that for well-posed, criteria with more directly specified operational anchors, three capable, similarly-aligned LLMs converge closely enough that an undifferentiated pipeline and a typed-abstention pipeline behave identically — typed abstention costs nothing there, but it also buys nothing. Pilot 2 shows that once criteria become normatively contested and a weaker or differently-trained model joins the panel, real disagreement appears, and the gate detects it without being told where to look — specifically on the criteria a domain expert would likely flag as most contestable (political acceptability, distributive fairness), not on the technical ones (road-space reallocation, gentrification risk stayed coherent even in Pilot 2). The practical implication is that typed abstention's value is concentrated where intuition would place it: contested, value-laden criteria and heterogeneous or lower-quality elicitation sources — a testable design rule, not a hedge.

Limitations. Neither pilot has a ground-truth or expert reference ranking, so conditional error and loss under declared abstention costs — the trade-off the source framework [4] is really about — could not be computed; coverage and the naive-vs-gated score comparison are the only metrics reported. Both pilots use equal criterion weights and the same 4×4 illustrative case; a real decision needs domain-informed weights and, ideally, a larger and more heterogeneous panel with $k > 2$. Pilot 2's criteria with greater normative and contextual dependence elicit normative opinions, not facts, so its CONFLICTING flags indicate an unreliable source, not a correct answer.

4.1. Neutrosophic reading of typed abstention

The neutrosophic interpretation is operationalized for the three withheld cells in Table 5. T is the share of usable models whose within-model pair does not cross τc; I is the share that does; and F is the between-model representative spread divided by the four-level scale width. These quantities preserve different evidence defects rather than averaging them into confidence. In this experiment only COHERENT and CONFLICTING states occurred; OPPOSED and INSUFFICIENT remain design states supported by synthetic tests, not empirically observed outcomes. Pilot 1 and Pilot 2 are not direct replications because the criteria, model panel, call count and surviving TOPSIS weights changed.

Table 8. Operational neutrosophic profile for the three withheld Pilot-2 cells.

Pilot-2 cell	T: coherent-model share	I: internal-conflict share	F: spread/4	Gate type
BRT–CULTFIT	0.75	0.25	0.25	CONFLICTING
LRT–CULTFIT	0.75	0.25	0.25	CONFLICTING
Minibus–DISP	0.50	0.50	0.25	CONFLICTING

5. Conclusions

We ported a source-preserving, typed-abstention decision policy into fuzzy MCDM pipelines where criteria are elicited from LLMs, implemented it fully, and ran it on two real, live pilots. Pilot 1 (96 calls) came back all-coherent, an honest null result; Pilot 2 (128 calls), redesigned with criteria of greater normative and contextual dependence and a weaker model, shows the gate firing on genuine disagreement and materially changing the evaluation matrix. Next steps: (1) obtain a domain-expert or literature-derived reference ranking to compute conditional error and loss, not just coverage; (2) confirm or replace the illustrative case with an existing transportation dataset; (3) decide real, non-equal criterion weights; (4) benchmark typed abstention against undifferentiated fuzzy TOPSIS/VIKOR on the full coverage/error/loss trade-off.

Funding: This research received no external funding.

Data Availability Statement: Code and data (prompts, raw LLM responses, computed results for both pilots) are archived in Documents/Qatar_Applications_2026/Paper_Aydin_TypedAbstention_MCDM/ and will be released in a public repository upon acceptance.

References

Aydin, N.; Samanlıoğlu, F.; Sert, Y.B.; et al. Performance Evaluation of Operators in the Telecommunication Industry. Int. J. Fuzzy Syst. 2025. https://doi.org/10.1007/s40815-025-02044-7

Aydin, N.; Carı, M.; Kara, B. A Review of Artificial Intelligence-Enhanced Fuzzy Multi-Criteria Decision-Making Approaches for Sustainable Transportation Planning. Comput. Mater. Contin. 2025. https://doi.org/10.32604/cmc.2025.067290

Leyva-Vázquez, M.; Smarandache, F. Breaking the Chains of Probability. arXiv:2605.24053, 2026.

Leyva, M. Source Preserving Multicriteria Decisions with Typed Abstention. SSRN 2026. https://ssrn.com/abstract=7441661

Cetinkale, Z.; Aydın, N. Health Care Logistics Network Design and Analysis on Pandemic Outbreaks: Insights From COVID-19. Transp. Res. Rec. 2022. https://doi.org/10.1177/03611981221099015

Leyva-Vázquez, M.; Smarandache, F. Institutional Prestige as Geographic Bias in LLMs. arXiv:2608.18107, 2026.

Grattafiori, A.; et al. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

Hamidieh, K.; Thost, V.; Gerych, W.; et al. Complementing Self-Consistency with Cross-Model Disagreement for Uncertainty Quantification. arXiv:2604.17112, 2026.

Liu, N. LLMs as a Jury: Cross-Model Consensus Can Outperform Process Reward Models for LLM Reasoning. arXiv:2607.10139, 2026.

Badshah, S.; Emami, A.; Sajjad, H. SCOPE: Selective Conformal Optimized Pairwise LLM Judging. arXiv:2602.13110, 2026.

Wang, Y.-J.; Lee, H.-S. Generalizing TOPSIS for Fuzzy Multiple-Criteria Group Decision-Making. Comput. Math. Appl. 2007. https://doi.org/10.1016/j.camwa.2006.08.037

Denœux, T. 40 Years of Dempster-Shafer Theory. Int. J. Approx. Reason. 2016. https://doi.org/10.1016/j.ijar.2016.07.010

Fei, L.; Xia, J.; Feng, Y.; et al. An ELECTRE-Based Multiple Criteria Decision-Making Method for Supplier Selection Using Dempster-Shafer Theory. IEEE Access 2019. https://doi.org/10.1109/access.2019.2924945

L. G. Lopezdominguez Rivas, S. D. Lopezdominguez Rivas, D. M. Lopezdominguez Rivas, and F. Parrales-Bravo, "How Much of a Multicriteria Ranking Is the Data and How Much Is the Arithmetic? A Neutrosophic Stress Test on Stakeholder Perception Data," accepted manuscript, CITIS 2026, paper 241, 2026.

M.-E. Puebla-Martínez, "HERA 2.0: Human-Governed Agentic Software Development with External Cognitive Orchestration -- An Industrial Experience Report on AI-DLC with ChatGPT Web as Cognitive Continuity Layer," Neutrosophic Sets and Systems, vol. 101, 2026, forthcoming.

González Vargas, Y.; Padrón Carrasco, D. When to refer instead of classify: Selective prediction with typed abstention in optometric screening. Journal of Evidential and Paraconsistent Reasoning 2026, 2(1). https://doi.org/10.5281/zenodo.22755549

Leyva-Vázquez, M.; Pérez-Teruel, K.; Febles-Estrada, A.; Gulín-González, J. Causal knowledge representation techniques: A case study in medical informatics. Revista Cubana de Información en Ciencias de la Salud 2013, 24(1), 73–83. https://www.medigraphic.com/pdfs/acimed/aci-2013/aci131f.pdf

Estupiñán Ricardo, J.; Leyva Vázquez, M.Y.; Álvarez Gómez, S.D.; Velázquez-Soto, O.E.; Rodríguez-Guzmán, A.A. The Application of Neutrosophy in Medical Sciences: A Narrative Literature Review. Revista Cubana de Información en Ciencias de la Salud 2023, 34, e2599.