SUPPLEMENTARY INFORMATION

Comparing Data-Derived and LLM-Elicited Indicator Weights for EU Circular Economy Assessment: An MCDA Study

Puren Veziroglu – Bicer[a*]

[a] Department of Agricultural Economics, Faculty of Agriculture, Çukurova University, 01330

Adana, Türkiye

[*] Corresponding author: Puren Veziroglu Bicer

pveziroglu@cu.edu.tr

Section S1

The final matrix included material footprint, resource productivity, municipal waste generated per capita, plastic packaging waste generated per capita, municipal waste recycling rate, packaging waste recycling rate, WEEE recycling rate, trade in recyclable raw materials, circular material use rate, employment in circular-economy sectors, circular-economy investment, consumption footprint and greenhouse-gas emissions from production.

Table S1 Circular economy indicators included in the analysis

Dimension	**Indicator**	**Eurostat Code**	**Unit of measure**	**Direction**	**Note**
Production & Consumption	Material footprint	cei_pc020	Tonnes per capita	Lower is better	-
Production & Consumption	Resource productivity	cei_pc030	Euro per kilogram	Higher is better	-
Production & Consumption	Municipal waste generated	cei_pc031	Kilograms per capita	Lower is better	-
Production & Consumption	Plastic packaging waste generated	cei_pc050	Kilograms per capita	Lower is better	-
Waste Management	Municipal waste recycling rate	cei_wm011	%	Higher is better	-
Waste Management	Packaging waste recycling rate	cei_wm020	Rate	Higher is better	-
Waste Management	WEEE recycling rate	cei_wm060	%	Higher is better	*
Secondary Raw Materials	Trade in recyclable raw materials	cei_srm020	kg per capita (converted from tonnes; Eurostat tps00001, 2022 population)	Higher is better	**
Secondary Raw Materials	Circular material use rate	cei_srm030	%	Higher is better	-
Competitiveness & Innovation	Persons employed in circular economy sectors	cei_cie011	Per 1,000 population (converted from FTE count; Eurostat tps00001, 2022 population)	Higher is better	**
Competitiveness & Innovation	Circular economy investment	cei_cie012	% of GDP	Higher is better	**
Global Sustainability & Resilience	Consumption footprint	cei_gsr010	PB	Lower is better	***
Global Sustainability & Resilience	GHG emissions from production activities	cei_gsr011	Kilograms CO2-equivalent per capita	Lower is better	-

*Romania and Sweden estimated via 3-year trailing moving average years 2019-2021

**Originally collected in absolute values and adjusted for population or GDP before the analysis.

*** The consumption-footprint indicator (`cei_gsr010`) was retrieved using the “Times the Planetary Boundary” unit and the single-weighted-score specification.

Section S2

S2. Missing-data treatment

The WEEE recycling-rate indicator contained missing 2022 observations for Romania and Sweden. Because 2022 and 2023 observations were unavailable for these countries, their missing values were estimated using the three-year mean of the available 2019–2021 observations. As a value-replacement sensitivity check, the imputed values were replaced with the respective 2021 observations and the CRITIC weights and TOPSIS rankings were recalculated. The resulting country ranking was identical, with Spearman's ρ equal to 1.000. A second check excluding the WEEE indicator entirely produced ρ = 0.9768 relative to the full model, with a mean absolute rank change of 1.11 positions.

Table S2 Number of EU Member States with available data by indicator and year

Indicator	2021	2022	2023	2024
cei_pc020	27	27	27	27
cei_pc030	27	27	27	27
cei_pc031	27	27	27	20
cei_pc050	26	27	23	n/a
cei_wm011	27	27	27	20
cei_wm020	26	27	23	n/a
cei_wm060	27	25	25	n/a
cei_srm020	27	27	27	27
cei_srm030	27	27	27	27
cei_cie011	27	27	27	n/a
cei_cie012	27	27	27	n/a
cei_gsr010	27	27	27	27
cei_gsr011	27	27	27	27

S2.1. Indicator transformations

For the per-capita transformations, the original values for persons employed in circular economy sectors and trade in recyclable raw materials were divided by the 2022 population obtained from Eurostat dataset *tps00001*. Converting absolute measures to population-adjusted values improves cross-country comparability and is consistent with established practice in composite indicator construction (Joint Research Centre, 2008). Circular economy investment was re-extracted as a percentage of GDP from another sub-series of the same Eurostat dataset, replacing the initially extracted values reported in million euros. Eurostat indicator data were downloaded on 5 August 2026. Population data used for country-size adjustments were obtained from the Eurostat `tps00001` dataset, downloaded on 9 August 2026. The reference year was 2022. Two indicators required country-size adjustments. Employment in circular-economy sectors was expressed per 1,000 inhabitants and trade in recyclable raw materials was expressed in kilograms per capita. Circular-economy investment was retained as a percentage of GDP. These transformations were applied before weighting and ranking.

Section S3

S3.1.AI Virtual Expert Panel: Full Elicitation Protocol and Prompt Texts

The elicitation used Claude (`claude-sonnet-5`) and ChatGPT (`gpt-5.6-terra`). Each persona–model combination was queried five times, resulting in 60 sessions: 10 dimension-level sessions and 50 indicator-level sessions. All sessions were conducted on 18 August 2026. Neither model was given the country-level decision matrix or the data-derived weights. The prompts asked the models to judge the relative importance of the criteria using Saaty's 1–9 pairwise-comparison scale. Each pairwise-comparison matrix was converted into a priority vector using the principal eigenvector method. Consistency was checked using Saaty's Consistency Ratio (CR), with CR < 0.10 used as the conventional threshold. Matrices with CR values above this threshold were not removed automatically; instead, they were kept and reported. The order of dimensions and indicators was fixed and remained the same across all repetitions. No randomization of presentation order was used. Therefore, possible order effects cannot be ruled out and are reported as a limitation. The collection scripts did not set a specific temperature value, so both model families were queried using their default sampling settings. ChatGPT was additionally queried with `reasoning_effort = medium`. No user-defined random seed was recorded. Since the LLMs did not receive country-level numerical data, this change did not directly affect the AHP calculations. However, the wording of an indicator may influence model judgments. For this reason, the difference between the prompt wording and the final data unit is reported as a protocol limitation. Each repetition was parsed into an AHP comparison matrix using a format-agnostic parsing procedure. All 60 sessions were parsed successfully. The resulting priority vectors were combined using an AIP-type aggregation procedure based on the elementwise geometric mean. The final weights were then normalized so that they summed to one. For the pooled panel, the five Claude and five ChatGPT priority vectors were treated as ten equally weighted vectors and combined directly using the geometric mean. The two model families were not averaged separately before pooling. Stability was examined at the local Stage-2 priority vector level before the Stage-2 weights were multiplied by the Stage-1dimension weights. Coefficients of variation were used to examine the stability of weight magnitudes. Pairwise rank correlations were also used when the number of indicators was large enough for a meaningful comparison. For dimensions with only two indicators, rank correlation provides little useful information, and the consistency ratio is structurally zero. These cases were therefore interpreted with caution. The LLM panel was used as a synthetic preference-elicitation tool and not as a replacement for human experts, stakeholders, or participatory modelling. Its purpose was to provide an additional and reproducible perspective on criterion importance and to show where data-derived weighting and LLM-based judgment weighting produce similar or different results.

S3.2. Model ID Overview

The panel comprised six personas: one neutral persona (Stage 1, cross-dimension comparisons) and five domain-expert personas (Stage 2, within-dimension comparisons, one per CEMF thematic dimension). Each persona was elicited from two model families (Claude, model claude-sonnet-5; ChatGPT, model gpt-5.6-terra) with k=5 independent repetitions per persona-model combination and randomized presentation order across repetitions, yielding 60 total elicitation sessions (10 in Stage 1, 50 in Stage 2). Both models were queried using default sampling parameters (temperature and top-p); Claude's API does not require or expose an explicit temperature parameter for this model tier and ChatGPT's API rejected a non-default temperature value, defaulting instead to its standard setting.

S3.3. System prompt

Stage 1: Cross-dimension comparison (neutral persona)

You are a sustainability policy analyst with broad expertise across circular economy domains. You are participating in an independent expert panel assessing the relative importance of five thematic dimensions used to evaluate the circular economy performance of EU member states, based on the EU Circular Economy Monitoring Framework (Eurostat). Your judgments will be combined with those of other independent experts. Answer based on your professional judgment of substantive importance -- not on statistical properties of the data, which you have not seen. Use Saaty's 1-9 scale for every comparison, and briefly justify each judgment in one sentence.

Task prompt

Compare the following five thematic dimensions in terms of their relative importance for assessing a country's overall circular economy performance:

(1) Production and Consumption

(2) Waste Management

(3) Secondary Raw Materials

(4) Competitiveness and Innovation

(5) Global Sustainability and Resilience

Provide all 10 pairwise judgments (1 vs 2, 1 vs 3, ... 4 vs 5) using the 1-9 scale. Present your answer as a table: [Dimension A], [Dimension B], [Score], [One-sentence justification].

Stage 2: Within-dimension comparison (five domain-expert personas)

All five Stage-2 personas share the same system-prompt template, with only the bracketed persona description varying:

You are [PERSONA]. You are participating in an independent expert panel assessing the relative importance of indicators used to evaluate the circular economy performance of EU member states, based on the EU Circular Economy Monitoring Framework (Eurostat). Your judgments will be combined with those of other independent experts. Answer based on your professional judgment of substantive importance -- not on statistical properties of the data, which you have not seen. Use Saaty's 1-9 scale for every comparison, and briefly justify each judgment in one sentence.

Persona descriptions (substituted for [PERSONA] above)

1. Production and Consumption: "a senior economist specializing in resource efficiency and material flow analysis in EU production and consumption systems"

2. Waste Management: "a waste management policy specialist with expertise in EU municipal and packaging waste recycling systems"

3. Secondary Raw Materials: "a circular supply chain researcher specializing in secondary raw materials markets and material recovery"

4. Competitiveness and Innovation: "an innovation economist specializing in the labor market and investment dynamics of the circular economy sector"

5. Global Sustainability and Resilience: "an environmental policy analyst specializing in consumption footprints, greenhouse gas accounting, and resource dependency"

Task prompt Production and Consumption (4 indicators, 6 pairs)

Within the "Production and Consumption" dimension, compare the following four indicators in terms of relative importance for assessing circular economy performance:

- Material footprint (raw material extraction per capita; lower is more circular)

- Resource productivity (economic output per unit of material used; higher is more circular)

- Municipal waste generation per capita (lower is more circular)

- Plastic packaging waste generation per capita (lower is more circular)

Provide all 6 pairwise judgments using the 1-9 scale, same format as before.

Task prompt Waste Management (3 indicators, 3 pairs)

Within the "Waste Management" dimension, compare the following three indicators:

- Municipal waste recycling rate (%)

- Packaging waste recycling rate (%)

- WEEE (electronic waste) recycling rate (%)

Provide all 3 pairwise judgments using the 1-9 scale.

Task prompt Secondary Raw Materials (2 indicators, 1 pair)

Within the "Secondary Raw Materials" dimension, compare:

- Trade in recyclable raw materials (kg per capita; higher assumed to reflect a more developed secondary materials market)

- Circular material use rate (% of material use from recycled sources; higher is more circular)

Provide the pairwise judgment using the 1-9 scale. Answer in exactly this format on its own line: "Trade in recyclable raw materials vs Circular material use rate: Score = X" (where X is a number or fraction like 1/3), followed by a one-sentence justification.

Task prompt Competitiveness and Innovation (2 indicators, 1 pair)

Within the "Competitiveness and Innovation" dimension, compare:

- Persons employed in circular economy sectors (per 1,000 population)

- Circular economy investment (% of GDP)

Provide the pairwise judgment using the 1-9 scale. Answer in exactly this format on its own line: "Persons employed in circular economy sectors vs Circular economy investment: Score = X" (where X is a number or fraction like 1/3), followed by a one-sentence justification.

Task prompt Global Sustainability and Resilience (2 indicators, 1 pair)

Within the "Global Sustainability and Resilience" dimension, compare:

- Consumption footprint (PB; lower is more sustainable)

- Greenhouse gas emissions from production (kg CO2-equivalent per capita; lower is more sustainable)

Provide the pairwise judgment using the 1-9 scale. Answer in exactly this format on its own line: "Consumption footprint vs Greenhouse gas emissions from production: Score = X" (where X is a number or fraction like 1/3), followed by a one-sentence justification.

S3.4. Shared protocol parameters

Model families and versions: Claude (claude-sonnet-5, Anthropic) and ChatGPT (gpt-5.6-terra, OpenAI).

Repetitions: k=5 independent repetitions per persona-model combination.

Presentation order: indicator/dimension order was randomized independently for each repetition.

Sampling configuration: default (non-zero) sampling was used for both model families; the temperature parameter is not user-configurable for either model at the API level under the settings used here, and default reasoning-effort settings (“medium” for ChatGPT) were applied. Consistency Ratio (CR) threshold: 0.10, following Saaty's

convention. Matrices exceeding this threshold were retained and reported rather than automatically excluded. Aggregation: for each persona, priority weights from all repetitions of both model families were pooled (5 Claude + 5 ChatGPT = 10 judgments) and combined via the geometric mean, rather than averaging two separately aggregated per-model weight vectors.

Response parsing: a format-agnostic extraction procedure was used to accommodate the five distinct response-formatting styles observed across sessions (tabular, numbered-header with separate judgment line, explicit score statement, bulleted list, and combined-cell tabular). All 60 sessions were parsed successfully with zero unmatched judgments.

Table S3 Large language model identifiers, API settings and data-collection details of the AI-panel elicitation.

Item	**Claude**	**ChatGPT**
Provider / access	Anthropic API (Python SDK)	OpenAI API, Chat Completions (Python SDK)
Model identifier passed to the API	claude-sonnet-5	gpt-5.6-terra
Dated model snapshot	Not specified (identifier as above)	Not specified (identifier as above)
Date of data collection	18 August 2026	18 August 2026
Temperature	Default sampling (not set)[a]	Default sampling (not set)[a]
Top-p	Not set	Not set
Random seed	Not set	Not set
Reasoning setting	API default	reasoning_effort = "medium"
Maximum output length	3,000 tokens	3,000 completion tokens
Prompts	Identical system and task prompts for both models (Supplementary Section S4)	Identical system and task prompts for both models (Supplementary Section S4)
Repetitions (k)	5 per persona	5 per persona
Number of elicitations	30 (1 neutral persona × 5 + 5 domain personas × 5)	30 (1 neutral persona × 5 + 5 domain personas × 5)
Order of presentation	Fixed; not randomized[b]	Fixed; not randomized[b]
Context between calls	Each call independent (no conversation history)	Each call independent (no conversation history)
Raw responses	Stored unaltered, one row per API call (claude_raw_responses.csv)	Stored unaltered, one row per API call (chatgpt_raw_responses.csv)

[a] A temperature of 0.7 was planned in the protocol, but both APIs rejected non-default temperature values at the time of data collection; the parameter was therefore removed from both collection scripts and both models were queried with their default sampling behaviour. The default values were not logged.

[b] The pre-registration specified randomization of the presentation order; the elicitation used a fixed order.

Section S3.5. Aggregation Procedure

This section specifies the exact aggregation procedure used to combine repeated elicitations into the final 13-indicator AI-panel weight vector, addressing the distinction between Aggregation of Individual Judgments (AIJ) and Aggregation of Individual Priorities (AIP) (Forman & Peniwati, 1998).

3.5.1. Repetition-level aggregation (AIP)

Aggregation within each model. For each persona–model combination, the five pairwise-comparison matrices were analyzed separately. Each matrix was converted into a priority vector using the principal eigenvector method, and its

Consistency Ratio (CR) was calculated. The matrices were not combined before this step. The five resulting priority vectors were then aggregated using the element-wise geometric mean and normalized to sum to one. Therefore, aggregation was performed at the priority-vector level rather than at the level of the original pairwise judgments. This procedure follows the Aggregation of Individual Priorities (AIP) approach. Pooling Claude and ChatGPT. For each persona, the five Claude and five ChatGPT priority vectors were treated as ten equally weighted elicitation outputs. These ten vectors were combined directly using the element-wise geometric mean and then normalized. Separate Claude and ChatGPT vectors were also calculated for model-level comparisons, but they were not averaged to produce the pooled vector. Combining the two AHP stages. Stage 1 and Stage 2 were aggregated separately. Stage 1 produced the weights of the five CEMF dimensions, while Stage 2 produced the local indicator weights within each dimension.

The final global weight of indicator *i* in dimension *d* was calculated as:

$$w_i = w_d \, x \, w_{i|d}$$

where w_d is the aggregated weight of dimension *d,* and $w_{i|d}$ is the aggregated local weight of indicator *i* within that dimension. The resulting 13 global indicator weights were normalized to sum to one. Because the two stages were aggregated separately, individual Stage 1 and Stage 2 repetitions were not matched. For example, Stage 1 repetition 3 was not paired with Stage 2 repetition 3.

The coefficient of variation (CV) and mean pairwise Spearman correlation reported in Section 4.5.1 were calculated from the Stage 2 local priority vectors before they were multiplied by the Stage 1 dimension weights. Therefore, the analysis assessed repetition stability within each dimension, but it did not directly assess the stability of the final 13 global indicator weights. A repetition-matched global stability analysis would require combining specific Stage 1 and Stage 2 repetitions to produce repetition-specific global weight vectors. This analysis was not conducted and is considered a limitation and an area for future research.

Section S4

Table S4a The ranking of the 27 EU member states

	Scores		Rank	
Country	**TOPSIS**	**MARCOS**	**TOPSIS**	**MARCOS**
Netherlands	0.6371	0.7009	**1**	1
Slovenia	0.5707	0.5901	**2**	3
Slovakia	0.5478	0.5514	**3**	9
Belgium	0.5465	0.5988	4	2
Italy	0.5449	0.5792	5	4
France	0.5387	0.5674	6	6
Latvia	0.5380	0.5518	7	8
Germany	0.5279	0.5529	8	7
Croatia	0.5164	0.5453	9	11
Estonia	0.5135	0.5459	10	10
Sweden	0.5056	0.5276	11	14
Austria	0.5003	0.5390	12	13
Spain	0.5002	0.5411	13	12
Luxembourg	0.4914	0.5704	14	5
Lithuania	0.4900	0.5014	15	15
Poland	0.4782	0.4914	16	17
Czechia	0.4781	0.4947	17	16
Hungary	0.4662	0.4755	18	19
Bulgaria	0.4594	0.4585	19	23
Romania	0.4522	0.4507	20	25
Portugal	0.4437	0.4692	21	20
Greece	0.4224	0.4450	22	26
Malta	0.4213	0.4659	23	21
Cyprus	0.4145	0.4522	24	24
Finland	0.4013	0.4283	**25**	27
Denmark	0.3955	0.4785	**26**	18
Ireland	0.3943	0.4598	**27**	22

Table S4b Agreement between TOPSIS and MARCOS country rankings under six weight sets (Panel A), and agreement between weight sets under each ranking method (Panel B)

Panel A. TOPSIS versus MARCOS within each weight set

Weight set	Spearman ρ	Mean |Δ rank|	Max |Δ rank|	Country with the largest difference (TOPSIS → MARCOS rank)	Top-5 overlap
CRITIC	0.909	2.30	9	Luxembourg (14 → 5)	4 of 5
Entropy	0.944	1.85	8	Cyprus (10 → 18)	3 of 5
MEREC	0.949	1.70	8	Ireland (4 → 12)	3 of 5
Pooled LLM	0.855	2.96	12	Luxembourg (19 → 7)	4 of 5
Claude	0.944	1.78	8	Estonia (18 → 10)	4 of 5
ChatGPT	0.695	4.67	15	Bulgaria (7 → 22)	3 of 5

Panel B. Spearman ρ between the rankings of two weight sets, by ranking method

Weight set 1	Weight set 2	TOPSIS	MARCOS
CRITIC	Entropy	0.372	0.756
CRITIC	MEREC	0.511	0.898
CRITIC	Pooled LLM	0.780	0.904
CRITIC	Claude	0.777	0.933
CRITIC	ChatGPT	0.681	0.863
Entropy	MEREC	0.923	0.938
Entropy	Pooled LLM	0.180	0.635
Entropy	Claude	0.412	0.706
Entropy	ChatGPT	-0.032	0.545
MEREC	Pooled LLM	0.400	0.822
MEREC	Claude	0.651	0.879
MEREC	ChatGPT	0.109	0.752
Pooled LLM	Claude	0.878	0.976
Pooled LLM	ChatGPT	0.861	0.978
Claude	ChatGPT	0.578	0.925

Note: Weights are held fixed within each comparison; only the ranking method changes. Rank 1 = best. Both methods used the same decision matrix. |Δ rank| is the absolute difference between the TOPSIS and MARCOS rank of a country. Top-5 overlap is the number of countries that appear in both top-five lists.

Table S5 Variation in criterion weights within each method (CRITIC, Entropy, and MEREC)

Method	Min	Max	Range	Std
CRITIC	0.0626	0.0888	0.0262	0.0086
Entropy	0.0030	0.3544	0.3514	0.0951
MEREC	0.0192	0.2065	0.1873	0.0526

Section S5

Table S6 Complete TOPSIS Scores and Rankings CRITIC vs. Pooled LLM Panel

Country	**Score**		**Rank**	
	CRITIC	Pooled LLM panel	CRITIC	Pooled LLM panel
Netherlands	0.6371	0.7380	**1**	**1**
Slovenia	0.5707	0.5701	**2**	9
Slovakia	0.5478	0.6327	**3**	5
Belgium	0.5465	0.6723	4	**2**
Italy	0.5449	0.6337	5	4
France	0.5387	0.6526	6	**3**
Latvia	0.5380	0.5424	7	11
Germany	0.5279	0.6296	8	6
Croatia	0.5164	0.5342	9	14
Estonia	0.5135	0.4806	10	22
Sweden	0.5056	0.5088	11	20
Austria	0.5003	0.5720	12	8
Spain	0.5002	0.5990	13	7
Luxembourg	0.4914	0.5104	14	19
Lithuania	0.4900	0.5065	15	21
Poland	0.4782	0.5357	16	13
Czechia	0.4781	0.5333	17	15
Hungary	0.4662	0.5502	18	10
Bulgaria	0.4594	0.5108	19	18
Romania	0.4522	0.4435	20	24
Portugal	0.4437	0.5386	21	12
Greece	0.4224	0.5195	22	17
Malta	0.4213	0.5276	23	16
Cyprus	0.4145	0.4290	24	**25**
Finland	0.4013	0.2915	**25**	**27**
Denmark	0.3955	0.4254	**26**	**26**
Ireland	0.3943	0.4486	**27**	23

Table S7 TOPSIS ranks of the 27 countries based on pooled, Claude-only and ChatGPT-only LLM weights (CRITIC for reference)

Country	Pooled	Claude	ChatGPT	Difference (Claude vs ChatGPT)	CRITIC
Netherlands	1	1	2	1	1
Belgium	2	2	4	2	4
France	3	4	3	1	6
Italy	4	5	8	3	5
Slovakia	5	9	1	8	3
Germany	6	3	10	7	8
Spain	7	8	9	1	13
Austria	8	7	18	11	12
Slovenia	9	10	5	5	2
Hungary	10	17	6	11	18
Latvia	11	11	14	3	7
Portugal	12	15	12	3	21
Poland	13	16	13	3	16
Croatia	14	14	15	1	9
Czechia	15	12	16	4	17
Malta	16	13	20	7	23
Greece	17	23	11	12	22
Bulgaria	18	24	7	17	19
Luxembourg	19	6	24	18	14
Sweden	20	20	19	1	11
Lithuania	21	19	21	2	15
Estonia	22	18	22	4	10
Ireland	23	21	25	4	27
Romania	24	26	17	9	20
Cyprus	25	25	23	2	24
Denmark	26	22	26	4	26
Finland	27	27	27	0	25

Table S8 Dimension-level weights from the neutral persona and the within-dimension share of consumption footprint, by model and for the pooled panel

Dimensions	Claude	ChatGPT	Pooled
Production and consumption	0.360	0.331	0.352
Waste management	0.205	0.173	0.192
Secondary raw materials	0.108	0.103	0.108
Competitiveness and innovation	0.184	0.104	0.141
Global sustainability and resilience	0.143	0.289	0.207
Share of consumption footprint within Global sustainability and resilience	0.600	0.750	0.680

Table S9 Comparison of indicator weights obtained from the pooled LLM expert panel and CRITIC

Code	Indicator	LLM expert panel (pooled)	CRITIC
cei_pc020	Material footprint	0.1775	0.0643
cei_gsr010	Consumption footprint	0.1408	0.0831
cei_pc030	Resource productivity	0.1114	0.0845
cei_wm011	Municipal waste recycling rate	0.1058	0.0791
cei_cie012	CE investment (% GDP)	0.1004	0.0830
cei_srm030	Circular material use rate	0.0898	0.0720
cei_gsr011	GHG emissions from production	0.0660	0.0888
cei_wm020	Packaging waste recycling rate	0.0543	0.0680
cei_pc031	Municipal waste generated	0.0417	0.0820
cei_cie011	Persons employed in CE sectors	0.0410	0.0789
cei_wm060	WEEE recycling rate	0.0320	0.0696
cei_pc050	Plastic packaging waste	0.0214	0.0626
cei_srm020	Trade in recyclable raw materials	0.0180	0.0841

5.1. Statistical measures and calculation settings

Coefficient of variation (CV). For each indicator, the CV was calculated across the five repetitions by dividing the population standard deviation by the mean weight. The population standard deviation was calculated in NumPy using its default setting (ddof = 0). The persona-level mean CV reported in the main text is the arithmetic mean of the CV values for all indicators evaluated by that persona. Pairwise Spearman correlation. For each persona-model combination, Spearman's rho was calculated for all possible pairs of the five repetition-level weight vectors. Five repetitions produced C(5,2) = 10 pairs. The reported mean pairwise rho is the arithmetic mean of the ten correlation coefficients. This measure was not reported for personas with only two indicators because a two-item ranking can only remain the same or be reversed and therefore provides limited information about rank stability.

Cosine similarity. Cosine similarity was calculated as the normalized dot product of the two 13-dimensional weight vectors: The calculation used scipy.spatial.distance.cosine.

Jensen–Shannon distance and divergence. Jensen–Shannon distance was calculated using scipy.spatial.distance.jensenshannon with a base-2 logarithm.

Spearman p-values. All reported Spearman p-values were two-sided, following SciPy's default setting (alternative = "two-sided"). No one-sided tests were used.

Table S10 Dimension and indicator level comparison of CRITIC and LLM weights

Dimension	CRITIC dimension	Pooled LLM dimension	Indicator	CRITIC global (flat)	CRITIC within-dimension	Pooled LLM global	Pooled LLM within-dimension
Production and Consumption	0.2934	0.3519	Material footprint	0.0643	0.2193	0.1775	0.5044
Production and Consumption	0.2934	0.3519	Resource productivity	0.0845	0.2879	0.1114	0.3164
Production and Consumption	0.2934	0.3519	Municipal waste generated per capita	0.0820	0.2796	0.0417	0.1185
Production and Consumption	0.2934	0.3519	Plastic packaging waste generated per capita	0.0626	0.2133	0.0214	0.0607
Waste Management	0.2167	0.1921	Municipal waste recycling rate	0.0791	0.3652	0.1058	0.5506
Waste Management	0.2167	0.1921	Packaging waste recycling rate	0.0680	0.3137	0.0543	0.2827
Waste Management	0.2167	0.1921	WEEE recycling rate	0.0696	0.3211	0.0320	0.1667
Secondary Raw Materials	0.1561	0.1078	Trade in recyclable raw materials	0.0841	0.5390	0.0180	0.1667
Secondary Raw Materials	0.1561	0.1078	Circular material use rate	0.0720	0.4610	0.0898	0.8333
Competitiveness and Innovation	0.1619	0.1413	Persons employed in CE sectors	0.0789	0.4874	0.0410	0.2899
Competitiveness and Innovation	0.1619	0.1413	CE investment	0.0830	0.5126	0.1004	0.7101
Global Sustainability and Resilience	0.1719	0.2069	Consumption footprint	0.0831	0.4835	0.1408	0.6808
Global Sustainability and Resilience	0.1719	0.2069	GHG emissions from production	0.0888	0.5165	0.0660	0.3192

Section S6

Table S11 Consistency Ratios Across Repetitions by Persona and Model

Persona	**Matrix**	**Model**	**Rep. 1**	**Rep. 2**	**Rep. 3**	**Rep. 4**	**Rep. 5**	**Mean**	**Max**
Neutral (dimension level)	5×5	Claude	0.0117	0.0030	0.0030	0.0022	0.0117	0.0063	0.0117
Neutral (dimension level)	5×5	ChatGPT	0.0259	0.0280	0.0204	0.0587	0.0704	0.0407	0.0704
Production and Consumption	4×4	Claude	0.0449	0.0395	0.0015	0.0121	0.0179	0.0232	0.0449
Production and Consumption	4×4	ChatGPT	0.0433	0.0433	0.0433	0.0433	0.0433	0.0433	0.0433
Waste Management	3×3	Claude	0.0079	0.0462	0.0158	0.0079	0.0079	0.0172	0.0462
Waste Management	3×3	ChatGPT	0.0079	0.0462	0.0079	0.0079	0.0079	0.0156	0.0462
Secondary Raw Materials	2×2	Claude	0*	0*	0*	0*	0*	0*	0*
Secondary Raw Materials	2×2	ChatGPT	0*	0*	0*	0*	0*	0*	0*
Competitiveness and Innovation	2×2	Claude	0*	0*	0*	0*	0*	0*	0*
Competitiveness and Innovation	2×2	ChatGPT	0*	0*	0*	0*	0*	0*	0*
Global Sustainability and Resilience	2×2	Claude	0*	0*	0*	0*	0*	0*	0*
Global Sustainability and Resilience	2×2	ChatGPT	0*	0*	0*	0*	0*	0*	0*

* For two-indicator personas the comparison matrix is 2×2, which is always perfectly consistent; CR is zero by construction and carries no information (grey rows). CR < 0.10 is the conventional acceptance threshold (Saaty, 1990). Across the 30 matrices with three or more items, CR ranged from 0.0015 to 0.0704 (median 0.017; Claude mean 0.016, ChatGPT mean 0.033). The two values above 0.05 (0.0587 and 0.0704) both occurred for ChatGPT in the neutral persona (repetitions 4 and 5)

Table. S12 Bootstrap rank uncertainty of the 27 countries (pooled, Claude, ChatGPT)

Country	Pooled weights, stratified resampling				Pooled weights, unstratified		Claude weights				ChatGPT weights			
	Rank	Median	90% interval	P(Top-5)	90% interval	P(Top-5)	Rank	Median	90% interval	P(Top-5)	Rank	Median	90% interval	P(Top-5)
Netherlands	1	1	1–1	100.0%	1–1	100.0%	1	1	1–1	100.0%	2	2	1–6	91.1%
Belgium	2	2	2–3	100.0%	2–4	99.9%	2	2	2–2	100.0%	4	4	3–10	62.4%
France	3	3	3–4	100.0%	3–4	100.0%	4	4	3–5	100.0%	3	3	3–6	93.1%
Italy	4	5	4–6	81.0%	4–6	84.9%	5	5	3–6	84.0%	8	8	5–17	24.4%
Slovakia	5	5	2–6	57.4%	2–6	53.7%	9	9	6–9	0.0%	1	1	1–2	99.9%
Germany	6	5	4–6	61.5%	4–6	58.6%	3	3	3–5	100.0%	10	10	6–16	0.1%
Spain	7	7	7–7	0.0%	7–8	0.0%	8	8	6–8	0.0%	9	9	7–13	0.0%
Austria	8	8	8–10	0.0%	8–12	0.0%	7	7	6–8	0.0%	18	18	17–20	0.0%
Slovenia	9	9	8–9	0.1%	7–10	2.8%	10	10	10–12	0.0%	5	5	3–8	69.5%
Hungary	10	10	9–12	0.0%	9–13	0.0%	17	17	13–19	0.0%	6	6	3–9	24.5%
Latvia	11	11	11–12	0.0%	11–13	0.0%	11	12	11–13	0.0%	14	13	11–14	0.0%
Portugal	12	12	11–17	0.0%	11–17	0.0%	15	15	14–17	0.0%	12	11	8–15	0.0%
Poland	13	13	12–16	0.0%	12–17	0.0%	16	17	14–18	0.0%	13	11	10–13	0.0%
Croatia	14	14	13–16	0.0%	13–16	0.0%	14	15	13–17	0.0%	15	15	14–16	0.0%
Czechia	15	15	12–17	0.0%	11–18	0.0%	12	12	10–14	0.0%	16	16	14–18	0.0%
Malta	16	16	13–19	0.0%	12–19	0.0%	13	12	11–15	0.0%	20	20	16–21	0.0%
Greece	17	17	16–19	0.0%	14–20	0.0%	23	23	20–23	0.0%	11	10	8–12	0.0%
Bulgaria	18	19	14–22	0.0%	11–22	0.1%	24	24	22–24	0.0%	7	7	2–18	31.9%
Luxembourg	19	19	10–22	0.0%	9–23	0.0%	6	6	4–9	16.0%	24	24	23–25	0.0%
Sweden	20	19	18–21	0.0%	17–21	0.0%	20	20	17–22	0.0%	19	18	14–20	0.0%
Lithuania	21	20	19–21	0.0%	18–21	0.0%	19	19	18–20	0.0%	21	20	19–21	0.0%
Estonia	22	22	22–23	0.0%	20–23	0.0%	18	18	14–21	0.0%	22	22	21–22	0.0%
Ireland	23	23	23–25	0.0%	22–25	0.0%	21	21	19–23	0.0%	25	25	23–25	0.0%
Romania	24	24	21–26	0.0%	21–26	0.0%	26	26	26–26	0.0%	17	17	7–23	3.0%
Cyprus	25	25	24–26	0.0%	24–26	0.0%	25	25	25–25	0.0%	23	23	23–24	0.0%
Denmark	26	26	24–26	0.0%	24–26	0.0%	22	22	20–24	0.0%	26	26	26–27	0.0%
Finland	27	27	27–27	0.0%	27–27	0.0%	27	27	27–27	0.0%	27	27	26–27	0.0%

Section S7 Normalization Sensitivity

Table S13 Normalization used by each method in the main analysis

Method	Normalization	Stage
CRITIC	Min–max (benefit/cost aware)	Weights
Entropy	Sum-based (share of column total)	Weights
MEREC	Linear ratio (benefit: min/x; cost: x/max)	Weights
AHP (LLM-elicited)	None (no data used)	Weights
TOPSIS	Min–max (default in all main runs)	Ranking
MARCOS	Own normalization relative to ideal / anti-ideal alternative	Ranking

Table S14 Agreement of TOPSIS country rankings under alternative normalizations with the default (min–max) ranking, by weight set.

Weight set	TOPSIS normalization	Spearman ρ vs. min–max	Largest rank shift (places)	Top-5 overlap	Bottom-5 overlap
CRITIC	Vector	0.654	19	3/5	1/5
CRITIC	Maximum	0.862	10	4/5	2/5
CRITIC	Sum	0.500	24	3/5	1/5
Entropy	Vector	0.962	5	5/5	5/5
Entropy	Maximum	0.995	3	5/5	5/5
Entropy	Sum	0.927	8	5/5	5/5
MEREC	Vector	0.961	5	4/5	4/5
MEREC	Maximum	0.998	1	5/5	5/5
MEREC	Sum	0.891	12	4/5	4/5
LLM pooled	Vector	0.850	12	4/5	4/5
LLM pooled	Maximum	0.886	12	4/5	4/5
LLM pooled	Sum	0.673	14	4/5	3/5
LLM Claude	Vector	0.909	9	4/5	4/5
LLM Claude	Maximum	0.931	8	4/5	5/5
LLM Claude	Sum	0.828	10	4/5	4/5
LLM ChatGPT	Vector	0.686	17	3/5	3/5
LLM ChatGPT	Maximum	0.788	14	3/5	4/5
LLM ChatGPT	Sum	0.558	17	3/5	2/5

Weights are held fixed; only the normalization in the TOPSIS ranking step changes. Bottom-5 = ranks 23–27. Top-5 and bottom-5 overlap are counted against the min–max ranking.

Table S15 CRITIC weights and the resulting TOPSIS ranking when CRITIC is computed under alternative normalizations.

Normalization used to compute CRITIC weights	Cosine similarity with default CRITIC weights	Spearman ρ of indicator weights vs. default	Spearman ρ of country ranking vs. default	Largest rank shift (places)	Cosine similarity with pooled LLM weights
Min–max (default)	1.000	1.000	1.000	0	0.849
Vector	0.902	0.330	0.654	19	0.738
Maximum	0.971	0.489	0.862	10	0.824
Sum	0.853	0.297	0.496	24	0.661

TOPSIS ranking uses min–max normalization in all rows; only the normalization inside CRITIC changes.

Table S16 Spearman ρ between country rankings from different weight sets, by TOPSIS normalization.

TOPSIS normalization	**LLM pooled vs. CRITIC**	**CRITIC vs. Entropy**	**CRITIC vs. MEREC**	**Entropy vs. MEREC**
Min–max (default)	0.780	0.372	0.511	0.923
Vector	0.625	0.795	0.899	0.913
Maximum	0.855	0.624	0.802	0.917
Sum	0.675	0.810	0.904	0.955
Vector (classic Hwang–Yoon)	0.625	0.795	0.899	0.913

Weights are held fixed. The classic Hwang–Yoon row (vector normalization of raw values) is identical to the Vector row.

Table S17 TOPSIS rank of each country under four normalizations, for CRITIC weights and for pooled LLM weights.

Country	**CRITIC min–max**	**CRITIC vector**	**CRITIC max**	**CRITIC sum**	**CRITIC range**	**LLM min–max**	**LLM vector**	**LLM max**	**LLM sum**	**LLM range**
Netherlands	1	1	1	1	0	1	1	1	1	0
Belgium	4	3	2	5	3	2	2	2	2	0
France	6	9	6	10	4	3	4	4	4	1
Italy	5	7	5	9	4	4	3	3	3	1
Slovakia	3	17	12	19	16	5	10	9	10	5
Germany	8	10	7	11	4	6	5	5	6	1
Spain	13	12	11	14	3	7	6	6	7	1
Austria	12	13	9	13	4	8	9	8	9	1
Slovenia	2	2	3	2	1	9	13	11	13	4
Hungary	18	24	22	24	6	10	14	15	19	9
Latvia	7	5	8	6	3	11	18	13	21	10
Portugal	21	20	21	21	1	12	15	16	20	8
Poland	16	22	18	20	6	13	19	17	22	9
Croatia	9	11	13	12	4	14	11	14	16	5
Czechia	17	21	16	23	7	15	12	12	17	5
Malta	23	16	17	16	7	16	8	10	8	8
Greece	22	23	26	22	4	17	16	21	14	7
Bulgaria	19	25	24	26	7	18	25	24	25	7
Luxembourg	14	4	4	4	10	19	7	7	5	14
Sweden	11	18	14	17	7	20	22	22	18	4
Lithuania	15	15	15	15	0	21	21	19	23	4
Estonia	10	6	10	8	4	22	20	18	11	11
Ireland	27	8	20	3	24	23	17	20	12	11
Romania	20	26	25	25	6	24	26	26	26	2
Cyprus	24	19	23	18	6	25	24	25	24	1
Denmark	26	14	19	7	19	26	23	23	15	11
Finland	25	27	27	27	2	27	27	27	27	0

Range = highest minus lowest rank across the four normalizations (places). Countries are sorted by their pooled-LLM rank under min–max normalization. 1 = best.

Table S18a. Agreement between the replication condition and the main results

Comparison	Cosine similarity	Spearman ρ (weights)	Max. absolute weight difference	Indicator with largest difference	Spearman ρ (TOPSIS ranks)	Max. rank shift	Countries with changed rank
replicate vs original claude	0.975	0.901	0.047	Consumption footprint	0.954	8	18
replicate vs original chatgpt	0.996	0.962	0.024	Circular material use rate	0.995	2	10
replicate vs original pooled	0.984	0.907	0.036	Consumption footprint	0.941	6	22
pooled, all 25 responses per model and set vs main results	0.990	0.940	0.029	cei_gsr010	0.955	6	22

Table S18b. Indicator weights in the main results and in the replication condition

Indicator	Main: pooled	Replication: pooled	Main: Claude	Replication: Claude	Main: ChatGPT	Replication: ChatGPT
Material footprint	0.1775	0.1856	0.1593	0.1696	0.1868	0.1864
Resource productivity	0.1114	0.1158	0.1351	0.1424	0.0867	0.0865
Municipal waste generated	0.0417	0.0441	0.0423	0.0460	0.0389	0.0388
Plastic packaging waste	0.0214	0.0209	0.0236	0.0219	0.0183	0.0182
Municipal waste recycling rate	0.1058	0.1145	0.1110	0.1237	0.0965	0.0984
Packaging waste recycling rate	0.0543	0.0536	0.0632	0.0686	0.0447	0.0389
WEEE recycling rate	0.0320	0.0313	0.0306	0.0372	0.0321	0.0245
Trade in recyclable raw materials	0.0180	0.0243	0.0181	0.0251	0.0172	0.0220
Circular material use rate	0.0898	0.1215	0.0903	0.1257	0.0859	0.1098
Persons employed in CE sectors	0.0410	0.0313	0.0613	0.0455	0.0261	0.0199
CE investment (% GDP)	0.1004	0.0759	0.1226	0.0911	0.0783	0.0586
Consumption footprint	0.1408	0.1046	0.0859	0.0393	0.2164	0.2238
GHG emissions from production	0.0660	0.0768	0.0567	0.0638	0.0721	0.0742

Table S18c. Pooled TOPSIS ranks in the main results and in the replication condition (shift: negative = better rank in the replication)

Country	Main rank	Replication rank	Shift
Belgium	2	2	0
Bulgaria	18	23	5
Czechia	15	11	-4
Denmark	26	25	-1
Germany	6	5	-1
Estonia	22	20	-2
Ireland	23	22	-1
Greece	17	19	2
Spain	7	7	0
France	3	4	1
Croatia	14	12	-2
Italy	4	3	-1
Cyprus	25	24	-1
Latvia	11	13	2
Lithuania	21	21	0
Luxembourg	19	15	-4
Hungary	10	14	4
Malta	16	10	-6
Netherlands	1	1	0
Austria	8	9	1
Poland	13	16	3
Portugal	12	18	6
Romania	24	26	2
Slovenia	9	8	-1
Slovakia	5	6	1
Finland	27	27	0
Sweden	20	17	-3

Table S18d. 90% bootstrap rank intervals (stratified by model, 5,000 resamples): main results (5 + 5 responses) and replication condition (20 + 20 responses)

Country	Main: rank	Main: 5th pct	Main: 95th pct	Main: P(top 5)	Main: P(bottom 5)	Repl.: rank	Repl.: 5th pct	Repl.: 95th pct	Repl.: P(top 5)	Repl.: P(bottom 5)
Netherlands	1	1	1	1	0	1	1	1	1	0
Belgium	2	2	3	1	0	2	2	3	1	0
France	3	3	4	1	0	4	4	4	1	0
Italy	4	4	6	0.81	0	3	2	3	1	0
Slovakia	5	2	6	0.57	0	6	6	7	0.00	0
Germany	6	4	6	0.61	0	5	5	5	1.00	0
Spain	7	7	7	0	0	7	6	7	0	0
Austria	8	8	10	0	0	9	8	9	0	0
Slovenia	9	8	9	0.00	0	8	8	9	0	0
Hungary	10	9	12	0	0	14	11	15	0	0
Latvia	11	11	12	0	0	13	11	14	0	0
Portugal	12	11	17	0	0	18	15	19	0	0
Poland	13	12	16	0	0	16	15	18	0	0
Croatia	14	13	16	0	0	12	11	14	0	0
Czechia	15	12	17	0	0	11	11	14	0	0
Malta	16	13	19	0	0	10	10	10	0	0
Greece	17	16	19	0	0	19	16	19	0	0
Bulgaria	18	14	22	0	0.02	23	22	23	0	0.69
Luxembourg	19	10	22	0	0.05	15	11	19	0	0
Sweden	20	18	21	0	0	17	15	19	0	0

Lithuania	21	19	21	0	0	21	20	21	0	0
Estonia	22	22	23	0	0.06	20	20	21	0	0
Ireland	23	23	25	0	0.97	22	22	23	0	0.31
Romania	24	21	26	0	0.90	26	26	26	0	1
Cyprus	25	24	26	0	1	24	24	25	0	1
Denmark	26	24	26	0	1	25	24	25	0	1
Finland	27	27	27	0	1	27	27	27	0	1

Note: The replication rank fell inside the original 90% interval for 16 of 27 countries (59%). The mean interval width was 3.00 positions in the main results and 1.74 with 20 responses per model.

Table S18e. Consistency ratio and identical responses (replication, shuffled and no-persona conditions)

model	group	**set**	**Mean CR**	**Max. CR**	**Distinct matrices (of 20)**	**Responses in most common matrix**
chatgpt	nopersona	competitiveness_innovation	0	0	4	12
chatgpt	nopersona	global_sustainability	0	0	1	20
chatgpt	nopersona	neutral	0.040	0.109	20	1
chatgpt	nopersona	production_consumption	0.043	0.052	4	17
chatgpt	nopersona	secondary_raw_materials	0	0	1	20
chatgpt	nopersona	waste_management	0.036	0.117	9	6
chatgpt	replicate	competitiveness_innovation	0	0	2	19
chatgpt	replicate	global_sustainability	0	0	3	18
chatgpt	replicate	neutral	0.041	0.109	19	2
chatgpt	replicate	production_consumption	0.043	0.043	1	20
chatgpt	replicate	secondary_raw_materials	0	0	1	20
chatgpt	replicate	waste_management	0.039	0.117	10	6
chatgpt	shuffled	competitiveness_innovation	0	0	2	17
chatgpt	shuffled	global_sustainability	0	0	2	17
chatgpt	shuffled	neutral	0.038	0.093	20	1
chatgpt	shuffled	production_consumption	0.071	0.277	11	9
chatgpt	shuffled	secondary_raw_materials	0	0	1	20
chatgpt	shuffled	waste_management	0.021	0.046	9	5
claude	nopersona	competitiveness_innovation	0	0	1	20
claude	nopersona	global_sustainability	0	0	2	18
claude	nopersona	neutral	0.019	0.049	19	2
claude	nopersona	production_consumption	0.030	0.050	7	9
claude	nopersona	secondary_raw_materials	0	0	1	20
claude	nopersona	waste_management	0.009	0.016	2	18
claude	replicate	competitiveness_innovation	0	0	1	20
claude	replicate	global_sustainability	0	0	3	16
claude	replicate	neutral	0.017	0.042	12	4
claude	replicate	production_consumption	0.036	0.084	12	4
claude	replicate	secondary_raw_materials	0	0	1	20
claude	replicate	waste_management	0.010	0.046	2	19
claude	shuffled	competitiveness_innovation	0	0	2	19
claude	shuffled	global_sustainability	0	0	1	20
claude	shuffled	neutral	0.023	0.050	19	2
claude	shuffled	production_consumption	0.033	0.061	17	2
claude	shuffled	secondary_raw_materials	0	0	1	20
claude	shuffled	waste_management	0.027	0.046	5	8

Table S19a. Agreement between shuffled-order and replication conditions

Comparison	Cosine similarity	Spearman ρ (weights)	Max. absolute weight difference	Indicator with largest difference	Spearman ρ (TOPSIS ranks)	Max. rank shift	Countries with changed rank
shuffled vs replicate claude	0.987	0.923	0.030	Consumption footprint	0.980	4	16
shuffled vs replicate chatgpt	0.995	0.934	0.029	Consumption footprint	0.974	4	17
shuffled vs replicate pooled	0.994	0.967	0.027	GHG emissions from production	0.984	4	13
shuffled vs ORIGINAL pooled	0.984	0.934	0.039	Circular material use rate	0.925	8	19
shuffled vs main results (pooled)	0.984	0.934	0.039	cei_srm030	0.925	8	19

Table S19b. Pooled weight differences (shuffled minus replication) with 90% bootstrap intervals

Indicator	Replication	Shuffled	Difference	5th pct	95th pct	Interval excludes 0
Material footprint	0.186	0.176	-0.009	-0.024	0.005	False
Resource productivity	0.116	0.128	0.013	0.002	0.023	True
Municipal waste generated	0.044	0.049	0.005	0.001	0.009	True
Plastic packaging waste	0.021	0.026	0.005	0.002	0.009	True
Municipal waste recycling rate	0.115	0.105	-0.009	-0.019	0.001	False
Packaging waste recycling rate	0.054	0.054	0.001	-0.005	0.007	False
WEEE recycling rate	0.031	0.027	-0.005	-0.009	-0.001	True
Trade in recyclable raw materials	0.024	0.026	0.002	-0.001	0.004	False
Circular material use rate	0.121	0.129	0.008	-0.007	0.022	False
Persons employed in CE sectors	0.031	0.033	0.002	-0.002	0.006	False

CE investment (% GDP)	0.076	0.077	0.001	-0.008	0.011	False
Consumption footprint	0.105	0.119	0.015	-0.001	0.031	False
GHG emissions from production	0.077	0.050	-0.027	-0.036	-0.018	True

Table S19c. Pooled TOPSIS ranks

Country	**Replication rank**	**Shuffled rank**	**Shift**
Belgium	2	2	0
Bulgaria	23	22	-1
Czechia	11	12	1
Denmark	25	24	-1
Germany	5	5	0
Estonia	20	20	0
Ireland	22	23	1
Greece	19	16	-3
Spain	7	7	0
France	4	4	0
Croatia	12	14	2
Italy	3	3	0
Cyprus	24	25	1
Latvia	13	17	4
Lithuania	21	21	0
Luxembourg	15	11	-4
Hungary	14	13	-1
Malta	10	10	0
Netherlands	1	1	0
Austria	9	9	0
Poland	16	15	-1
Portugal	18	19	1
Romania	26	26	0
Slovenia	8	8	0
Slovakia	6	6	0
Finland	27	27	0
Sweden	17	18	1

Note: The 90% interval of the weight difference excluded zero for five of 13 indicators; the largest difference was 0.027 (GHG emissions from production). Thirteen countries changed position, by at most four places. Two random halves of the replication responses (10 + 10 per model) gave a median Spearman ρ of 0.984 for the TOPSIS ranks (5th percentile 0.965).

Table S20a. Agreement between no-persona and replication conditions

Comparison	Cosine similarity	Spearman ρ (weights)	Max. absolute weight difference	Indicator with largest difference	Spearman ρ (TOPSIS ranks)	Max. rank shift	Countries with changed rank
nopersona vs replicate claude	0.983	0.879	0.033	CE investment (% GDP)	0.971	5	17
nopersona vs replicate chatgpt	0.976	0.973	0.067	Consumption footprint	0.872	11	25
nopersona vs replicate pooled	0.997	0.995	0.015	CE investment (% GDP)	0.995	2	13
nopersona vs ORIGINAL pooled	0.977	0.901	0.043	Circular material use rate	0.922	8	22
nopersona vs main results (pooled)	0.977	0.901	0.042	cei_srm030	0.922	8	22

Table S20b. Pooled weight differences (no persona minus replication) with 90% bootstrap intervals

Indicator	Replication	No persona	Difference	5th pct	95th pct	Interval excludes 0
Material footprint	0.186	0.179	-0.007	-0.016	0.003	False
Resource productivity	0.116	0.116	-0.000	-0.007	0.006	False
Municipal waste generated	0.044	0.046	0.002	-0.001	0.005	False
Plastic packaging waste	0.021	0.021	0.000	-0.001	0.002	False
Municipal waste recycling rate	0.115	0.121	0.007	-0.001	0.015	False
Packaging waste recycling rate	0.054	0.053	-0.000	-0.005	0.004	False
WEEE recycling rate	0.031	0.038	0.007	0.003	0.011	True
Trade in recyclable raw materials	0.024	0.026	0.002	-0.000	0.005	False
Circular material use rate	0.121	0.132	0.011	-0.001	0.023	False
Persons employed in CE sectors	0.031	0.029	-0.002	-0.005	0.002	False
CE investment (% GDP)	0.076	0.061	-0.015	-0.022	-0.007	True

Consumption footprint	0.105	0.110	0.005	-0.010	0.022	False
GHG emissions from production	0.077	0.066	-0.011	-0.022	-0.000	True

Table S20c. Pooled TOPSIS ranks

Country	**Replication rank**	**No-persona rank**	**Shift**
Belgium	2	3	1
Bulgaria	23	23	0
Czechia	11	11	0
Denmark	25	24	-1
Germany	5	5	0
Estonia	20	19	-1
Ireland	22	22	0
Greece	19	18	-1
Spain	7	7	0
France	4	4	0
Croatia	12	13	1
Italy	3	2	-1
Cyprus	24	25	1
Latvia	13	12	-1
Lithuania	21	21	0
Luxembourg	15	14	-1
Hungary	14	15	1
Malta	10	10	0
Netherlands	1	1	0
Austria	9	9	0
Poland	16	17	1
Portugal	18	20	2
Romania	26	26	0
Slovenia	8	8	0
Slovakia	6	6	0
Finland	27	27	0
Sweden	17	16	-1

Note: Three indicator weights changed (interval excluding zero), by at most 0.015. Rankings based on ChatGPT alone changed more when the persona instruction was removed (Spearman $\rho = 0.872$, 25 countries changed position) than rankings based on Claude alone ($\rho = 0.971$).

Table S21a. Indicator weights (global weights as reported; "n/a" = indicator not in the published set)

Indicator	Pooled LLM	Claude	ChatGPT	CRITIC	Entropy	MEREC	Expert [70]	Expert [71]
Material footprint	0.1775	0.1593	0.1868	0.0643	0.0428	0.0712	n/a	0.0115
Resource productivity	0.1114	0.1351	0.0867	0.0845	0.1263	0.1127	0.0330	0.0093
Municipal waste generated	0.0417	0.0423	0.0389	0.0820	0.0181	0.0328	0.0340	0.0183
Plastic packaging waste	0.0214	0.0236	0.0183	0.0626	0.0210	0.0524	n/a	0.0210
Municipal waste recycling rate	0.1058	0.1110	0.0965	0.0791	0.0539	0.0807	0.0577	0.0274
Packaging waste recycling rate	0.0543	0.0632	0.0447	0.0680	0.0131	0.0484	0.0272	0.0240
WEEE recycling rate	0.0320	0.0306	0.0321	0.0696	0.0030	0.0255	0.0528	0.0588
Trade in recyclable raw materials	0.0180	0.0181	0.0172	0.0841	0.3544	0.2065	0.0647	0.1725
Circular material use rate	0.0898	0.0903	0.0859	0.0720	0.1397	0.1252	0.0479	0.0082
Persons employed in CE sectors	0.0410	0.0613	0.0261	0.0789	0.0951	0.0533	0.1131	0.1111
CE investment (% GDP)	0.1004	0.1226	0.0783	0.0830	0.0928	0.1231	0.1530	0.0638
Consumption footprint	0.1408	0.0859	0.2164	0.0831	0.0069	0.0192	0.1184	0.0474
GHG emissions from production	0.0660	0.0567	0.0721	0.0888	0.0329	0.0491	0.1222	0.0498

Table S21b. Agreement with the published expert weights (matched indicators, rescaled to sum to one)

Reference	Weight set	Spearman's ρ	Top-3 shared (of 3)	Total weight shift	Same top indicator within dimension
[71] (13 matched)	LLM pooled	-0.46	0	0.487	0/5

[71] (13 matched)	Claude	-0.45	1	0.473	0/5
[71] (13 matched)	ChatGPT	-0.47	0	0.522	0/5
[71] (13 matched)	CRITIC	0.29	1	0.337	2/5
[71] (13 matched)	Entropy	0.01	1	0.35	3/5
[71] (13 matched)	MEREC	-0.03	2	0.352	2/5
[70] (11 matched)	LLM pooled	0.1	1	0.293	2/5
[70] (11 matched)	Claude	-0.03	1	0.298	2/5
[70] (11 matched)	ChatGPT	0.05	1	0.33	2/5
[70] (11 matched)	CRITIC	0.45	1	0.233	4/5
[70] (11 matched)	Entropy	0.04	0	0.486	3/5
[70] (11 matched)	MEREC	0.07	1	0.375	4/5

Section S8 Preregistration and deviations from the initial plan

The study was preregistered on the Open Science Framework on 10 August 2026, before the LLM-panel elicitation and the main analyses. The analyses were conducted on 17 and 18 August 2026, and the LLM-panel elicitation was conducted on 18 August 2026. The research questions specified during the initial study design and preregistration were later revised. The exact model specifications, software versions, collection scripts and analysis settings are also reported in the supplementary materials. These changes are reported to ensure transparency and to distinguish the initial registered plan from the final study design.

References

Joint Research Centre, 2008. Handbook on constructing composite indicators: methodology and user guide. OECD publishing.

Forman, E., & Peniwati, K. (1998). Aggregating individual judgments and priorities with the analytic hierarchy process. *European journal of operational research*, *108*(1), 165-169.