An Experimental Evaluation of Large Language Models in Supporting the DEX Multi-Criteria Decision-Making Process

Marko Bohanec*[1], Uroš Rajkovič[2], Vladislav Rajkovič[2]
[1]Jožef Stefan Institute, Department of Knowledge Technologies, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
[2]University of Maribor, Faculty of Organizational Sciences, Kidričeva cesta 55 A, SI-4000 Kranj, Slovenia
E-mail: marko.bohanec@ijs.si, uros.rajkovic@um.si, vladislav.rajkovic@gmail.com
*Corresponding author

Keywords: Multi-criteria decision-making, decision analysis, large language models, method DEX

Received: December 1, 2025

We experimentally assessed the capabilities of two mainstream artificial intelligence chatbots, ChatGPT and DeepSeek, to support the multi-criteria decision-making process. Specifically, we focused on using the method DEX (Decision EXpert) and investigated their performance in all stages of DEX model development and utilization. The results indicate that these tools may substantially contribute in the difficult stages of collecting and structuring decision criteria, and collecting data about decision alternatives. However, at the current stage of development, the support for the whole multi-criteria decision-making process is still lacking, mainly due to occasionally inconsistent and erroneous execution of methodological steps. To leverage the strengths of both approaches, we also propose a hybrid workflow for DEX model development that begins in the LLM and continues in the specialized DEXiWin software.

Povzetek: Eksperimentalno smo ocenili zmožnosti uveljavljenih klepetalnih umetnih inteligenc, ChatGPT in DeepSeek, pri podpori večkriterijskega odločanja z uporabo metode DEX (Decision EXpert). Preučili smo njuno učinkovitost v vseh fazah razvoja in uporabe modela DEX. Ugotovili smo, da orodji učinkovito podpirata zbiranje in strukturiranje kriterijev ter podatkov o alternativah, vendar je njuna podpora celotnemu procesu še omejena zaradi nedoslednega in občasno napačnega izvajanja metodoloških korakov. Za boljše rezultate predlagamo hibridni delotok, ki združuje začetno uporabo LLM in nadaljevanje razvoja modela v specializiranem programu DEXiWin.

1 Introduction

Multi-criteria decision-making (MCDM) [13] is an established approach to support decision-making in situations where it is necessary to consider multiple interrelated, and possibly conflicting criteria, and select the best solution based on the available alternatives and the preferences of the decision-maker. Traditionally, such models are developed in collaboration with decision makers and domain experts, who define the criteria, acquire decision makers' preferences and formulate the corresponding evaluation rules. The model-development process is demanding, as it includes structuring the problem, formulating all the necessary model components (such as decision preferences or rules) for evaluating decision alternatives, and analyzing the results.

With the development and success of generative artificial intelligence, especially large language models (LLMs) [12], the question arises as to how these models can support or perhaps partially automate decision-making processes. To this end, we explored the capabilities of recent mainstream LLM-based chatbots, specifically ChatGPT and DeepSeek, for supporting the MCDM process. We focused on using the method DEX (Decision EXpert) [5], with which we have extensive experience, spanning multiple decades [4], in the roles of decision makers, decision analysts, and teachers. DEX is a full-aggregation [7] multi-criteria decision modelling method, which proceeds by developing an explicit decision model. DEX uses qualitative (symbolic) variables to represent decision criteria, and decision rules to represent decision makers' preferences. Variables (attributes) are structured hierarchically, representing the decomposition of the decision problem into smaller, easier to handle subproblems. Traditionally, DEX models are developed using specialized software such as DEXiWin [6], which allows the users (decision makers, domain experts, decision analysts) to interactively construct a DEX model and use it to evaluate and analyze decision alternatives.

This study is of exploratory nature. We ran ChatGPT and DeepSeek multiple times over the last six months, either individually, as a group or in classrooms with students. Typically, we first formulated some hypothetical decision problem and then guided the chatbot through the following main stages of the MCDM process:

A. Model development stages:

1. Acquiring criteria
2. Definition of attributes (variables representing criteria)
3. Structuring attributes
4. Preference modeling (formulating decision rules)

B. Model utilization stages:

5. Definition of decision alternatives
6. Evaluation of alternatives
7. Explaining the results of evaluation
8. Analysis of alternatives

In doing this, we observed the responses generated by the LLMs and assessed them from the viewpoint of skilled decision analysts. The main goal was not to solve specific real-life decision problems, but to identify LLMs' strengths and weaknesses that may substantially affect the MCDM process.

Despite focusing on DEX, many of our findings are also applicable to other hierarchical full-aggregation MCDM methods [7][13], such as AHP, MAUT/MAVT, and MACBETH; they follow the same methodological stages, but represent model components differently, for instance with numeric variables and weight-based aggregation functions.

In the following sections, we first present related work on LLM for MCDM. We then examine each of the aforementioned MCDM stages, detailing our experience with them. Specifically, we illustrate the process with answers generated by ChatGPT-o3 and DeepSeek-V3, LLM versions that were available around May and June 2025. Basic concepts of DEX related to each stage are explained along the way. Finally, we propose a hybrid LLM-DEX workflow for developing decision models.

We considered a hypothetical personal decision problem of buying an electric-powered vehicle (EV). We assumed the role of an "ordinary" decision maker conducting a session with an LLM. The chatbots were run in parallel, using similar prompts. No API (Application Programming Interface) or other technique was used to repeat prompts, stabilize responses and/or assess the effects of LLM's parameters, such as "temperature". We assessed results subjectively from the standpoint of expert decision analysts. Illustrative examples presented below were obtained from two parallel chatbot sessions conducted on June 6, 2025, using similar prompts. Our assessments and comments are somewhat broader, based on some other use-cases, which are not presented here.

2 Related work

Despite the growing popularity of LLMs in recent scientific literature, their coverage within the context of MCDM is still limited. Several authors have suggested incorporating LLMs in the MCDM process. Wang et al. [17] proposed a framework using LLMs in the model preparation and evaluation stages, comparing it with the MCDM methods AHP (Analytic Hierarchy Process) and FCE (Fuzzy Comprehensive Evaluation). Similarly, Svoboda and Lande [16] proposed a decision analysis framework for cybersecurity that combines AHP with the GPT-4 LLM. The same LLM is used in software 1000minds (https://www.1000minds.com/), which incorporates an AI assistant to enhance user interaction by suggesting decision criteria and alternatives according to the MCDM method PAPRIKA (Potentially All Pairwise RanKings of all possible Alternatives). Yuan et al. [19] recognized and studied the potential of LLMs to extract and suggest candidate criteria from documents and stakeholder text.

Table 1: EV cars: Criteria suggested by DeepSeek

1. Range & Battery Efficiency
- **Real-world range** (WLTP or EPA estimates) – Aim for at least 300 km in mixed conditions.
- **Battery size** (kWh) – Larger batteries offer more range but may increase cost.
- **Energy consumption** (kWh/100 km) – Lower is better (e.g., 14–18 kWh/100 km).

2. Charging Speed & Compatibility
- **DC fast-charging capability** (e.g., 50 kW, 100 kW, or higher).
- **AC charging speed** (e.g., 7.4 kW, 11 kW, or 22 kW for home/work charging).
- **Charging port type** (CCS Combo is standard in Europe).

3. Performance & Driving Experience
- **Motor power** (kW/HP) – Affects acceleration and highway performance.
- **Drivetrain** (FWD, RWD, AWD) – AWD is better for Slovenia's varied weather.
- **Top speed & acceleration** (0–100 km/h time).

4. Practicality & Usability
- **Cargo space** (boot capacity, frunk if available).
- **Seating capacity** (4–5 seats, some EVs have limited rear space).
- **Ground clearance** (important for rural roads or winter conditions).

5. Price & Incentives
- **Purchase price** (must fit within €35,000).
- **Slovenian EV incentives** (check for subsidies, tax breaks, or benefits like free parking).
- **Insurance & maintenance costs** (EVs are generally cheaper to maintain).

6. Technology & Features
- **Infotainment system** (touchscreen, Apple CarPlay/Android Auto).
- **Driver assistance** (adaptive cruise, lane-keeping, parking sensors).
- **Heat pump availability** (improves winter range efficiency).

7. Availability & After-Sales Support
- **Local dealership & service network** in Slovenia.
- **Warranty coverage** (battery warranty should be 8+ years or 160,000+ km).

8. Resale Value & Future-Proofing
- **Brand reputation** (Tesla, Hyundai, Kia, VW, Renault, etc.).
- **Software updates** (OTA updates improve longevity).

Even fewer publications address the intersection of LLMs and DEX. Radovanović et al. [14] used LLMs to learn the structure and some components of DEX models from data, reporting that the approach was both effective (it identified appropriate criteria) and efficient (it reduced training time). Bohanec et al. [8] made a preliminary analysis of using LLM with DEX, and Bohanec [7] investigated the use of LLMs to create, modify and interpret DEX decision tables, reporting mixed results: while LLMs can interpret user inputs and supply substantial MCDM knowledge "out of the box," producing convincing explanations, they may also ignore the decision maker's stated preferences, implicitly shift context, and generate incorrect explanations.

3 Acquiring criteria

A MCDM model includes multiple criteria that capture essential aspects of decision alternatives. These criteria are used to evaluate and compare the alternatives in a structured way. Defining criteria usually requires a good knowledge of the decision problem and the decision maker's goals. This step is usually one of the most difficult and may require consulting domain experts and/or relevant literature.

We asked the chatbots: *I am considering buying a new EV vehicle for up to 35000 EUR in Slovenia. Suggest criteria for evaluating such cars.*

Both ChatGPT and DeepSeek came out with an extensive and structured list of criteria. Table 1 shows the criteria suggested by DeepSeek. ChatGPT's suggestions were similar, though they employed slightly different high-level categories, which also incorporated measurement units associated with each criterion.

Let us immediately say that we consider this the most important single contribution of LLMs to MCDM modeling. We are not aware of any previous method that would allow identifying and structuring decision criteria in such a depth and detail in literally just a minute. Of course, for "serious" applications getting such a list does not take the burden off the user, who is still responsible for verifying the suggestions and checking the criteria for relevance and correctness. Nevertheless, this is a valuable starting point that can save days or even weeks of work.

This stage does not depend on the MCDM method used, so other methods may benefit from using LLMs equally well. This is particularly true for hierarchical methods, which are designed to handle a large number of criteria organized in a multi-level hierarchical structure. Examples of such methods include AHP (Analytic Hierarchy Process [15]) and MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique [1]).

4 Definition of attributes

In this stage, the task is to define variables, called attributes, that represent criteria in a MCDM model. As most MCDM methods use numeric attributes, this stage is specific to DEX, which uses qualitative attributes. Therefore, this and the following stages require LLMs to "understand" the method used. While DEX is less widely known than methods such as AHP, it is nonetheless used and valued in various applications. Anyway, we were somewhat surprised to find out that all consulted LLMs were already familiar with DEX and reasonably capable of following its main methodological steps. In some steps, however, we had to specify additional requirements to obtain proper DEX model components.

Generally, defining qualitative value scales of attributes is not too difficult for LLMs. Asking *Suggest preferentially ordered value scales [for some attribute(s)]* typically gives good suggestions, for example (DeepSeek):

Purchase Price: High (>€34k) → Medium (€30k–34k) → Low (<€30k)
Government Incentives: None → Moderate (€1k–3k) → High (>€3k)
Insurance & Maintenance: Expensive → Moderate → Cheap

Interestingly, in our first attempts at using LLMs, they were not fully aware of DEX specifics, such as that value scales should have a small number of values, preferentially ordered from "bad" to "good" [5]. After explicitly articulating the rules and correcting LLMs a few times, they "remembered" and now suggest properly formulated value scales most of the times.

5 Structuring attributes

The next step after acquiring the criteria and defining attributes is to structure attributes into a hierarchy (most often an ordinary tree). Following the previous stages, which already resulted in a well-organized criteria structure and proper definition of attributes and their scales, this stage appears quite easy for LLMs. Figure 1 shows a full DEX model structure as suggested by ChatGPT without giving any further instructions.

In comparison, the DeepSeek's suggestion (not shown here) was somewhat inferior. Also, it was generated only after we had instructed it that DEX requires "narrow" trees with only two to three descendants for each aggregate attribute. Anyway, we consider both structures appropriate and comparable to the achievements of the best university-level students. The contributions of LLMs to DEX modeling up to this stage are substantial.

6 Decision preferences and decision rules

In DEX, decision maker's preferences are represented by decision rules, organized in decision tables. For each aggregate attribute (an internal variable in the model, which depends on subordinate variables), the user is asked to investigate all combinations of lower-level attribute values and assess the corresponding outcomes. Depending on the number of aggregate attributes, this can be a laborious task, though it is typically manageable using specialized software like DEXiWin. LLMs are also capable of suggesting perfectly valid decision tables, as illustrated in Table 2, which suggests the values of *Purchase-Cost* depending on *Net-Price-After-Subsidy* and *Registration-Fee*.

From the DEX perspective, it is important to remark that Table 2 is complete (addressing all possible combinations of input values) and preferentially consistent (better input values result in better or equal outputs). Initially, these requirements were not obvious to

```
EV-SUITABILITY                          {excellent, good, adequate, poor}
├─ COST-OWNERSHIP                       {excellent, good, adequate, poor}
│  ├─ ECONOMY                           {very-low-cost, low-cost, medium-cost, high-cost}
│  │  ├─ Purchase-Cost                  {very-low, low, medium, high}
│  │  │  ├─ Net-Price-After-Subsidy     {≤25 k, 25–30 k, 30–33 k, 33–35 k, >35 k}
│  │  │  └─ Registration-Fee %          {0.5, 1, 1–2, >2}
│  │  └─ Running-Cost                   {very-low, low, medium, high}
│  │     ├─ Electricity €/100 km        {≤3, 3–4, 4–5, >5}
│  │     └─ Service+Insurance €/yr      {≤600, 600–800, 800–1 000, >1 000}
│  ├─ DEPRECIATION-RISK                 {very-low, low, medium, high}
│  │  ├─ 5-yr-Residual-Value %          {≥65, 60–64, 55–59, <55}
│  │  └─ Brand-Resale-Index             {strong, above-avg, average, weak}
│  └─ DURABILITY                        {excellent, good, adequate, poor}
│     ├─ Vehicle-Warranty yrs           {≥7, 5–6, 3–4, <3}
│     ├─ Battery-Warranty yrs/%         {≥8/70, 8/60, 5–7/60, <5/60}
│     └─ Service-Interval km            {≥30 000, 20–29 000, 15–19 000, <15 000}
├─ CAPABILITY                           {excellent, good, adequate, poor}
│  ├─ RANGE-&-CHARGING                  {excellent, good, adequate, poor}
│  │  ├─ RANGE                          {very-high, high, medium, low, very-low}
│  │  │  ├─ WLTP-Range km               {≥500, 400–499, 300–399, 200–299, <200}
│  │  │  └─ Real-Efficiency kWh/100     {≤15, 15–17, 17–19, >19}
│  │  └─ CHARGING                       {excellent, good, adequate, poor}
│  │     ├─ DC-Peak kW                  {≥150, 100–149, 50–99, <50}
│  │     ├─ 10-80 %-Time min            {≤25, 26–35, 36–45, >45}
│  │     └─ AC-On-board kW              {≥11 3-ph, 11 1-ph, 7.4, <7}
│  ├─ PRACTICALITY                      {excellent, good, adequate, poor}
│  │  ├─ Cabin-&-Cargo                  {excellent, good, adequate, poor}
│  │  │  ├─ Boot-Volume L               {≥450, 400–449, 330–399, <330}
│  │  │  └─ Rear-Legroom mm             {≥900, 850–899, 800–849, <800}
│  │  └─ Versatility                    {excellent, good, adequate, poor}
│  │     ├─ Towing-Capacity kg          {≥1 000, 750–999, 500–749, <500}
│  │     └─ Roof-Load kg                {≥75, 50–74, <50}
│  └─ PERFORMANCE                       {excellent, good, adequate, poor}
│     ├─ 0-100 km/h s                   {≤7.5, 7.6–9.0, 9.1–11, >11}
│     ├─ Top-Speed km/h                 {≥180, 160–179, 140–159, <140}
│     └─ Drive-Layout                   {AWD, RWD, FWD}
└─ RESPONSIBILITY                       {excellent, good, adequate, poor}
   ├─ SAFETY-&-TECH                     {excellent, good, adequate, poor}
   │  ├─ SAFETY                         {excellent, good, adequate, poor}
   │  │  ├─ Euro-NCAP-Stars             {5★, 4★, 3★, ≤2★}
   │  │  └─ ADAS-Completeness #         {≥7, 5–6, 3–4, <3}
   │  └─ TECH-FEATURES                  {excellent, good, adequate, poor}
   │     ├─ Heat-Pump                   {yes, optional, no}
   │     ├─ V2L-Power kW                {≥3.6, 1.5–3.5, none}
   │     └─ OTA+Nav-Chargers            {full, partial, none}
   └─ SUSTAINABILITY                    {excellent, good, adequate, poor}
      ├─ Battery-Origin-Transparency    {EU+recycle, non-EU+recycle, non-EU, unknown}
      ├─ Recycled-Materials %           {≥25, 15–24, 5–14, <5}
      └─ Life-cycle CO2 t               {≤15, 16–18, 19–22, >22}
```

Figure 1: DEX model structure generated by ChatGPT

Table 2: Decision table for Purchase-Cost (ChatGPT)

Net-Price-After-Subsidy ▼ / Registration-Fee ▶	**very-low (0 .5 %)**	**low (1 %)**	**medium (1–2 %)**	**high (> 2 %)**
very-low (≤ 25 k €)	very-low	low	low	medium
low (25–30 k €)	low	low	medium	medium
medium (30–33 k €)	medium	medium	medium	high
high (33–35 k €)	high	high	high	high

LLMs, and we had to request them explicitly. After further use, LLMs now generate appropriate rules by themselves.

However, several critical issues must be emphasized. First, decision making is a subjective process, and decisions should reflect individual or group decision-makers' preferences. Using LLMs, we only get preferences, generalized from many sources that were used to train them. In order to impose our own preferences, we should (1) carefully check LLMs' suggestions and (2) request modifications, if necessary. According to our experience, this is possible by formulating prompts such as "*increase the importance of some attribute*" or "*alternatives exceeding some value of some attribute are unacceptable*". Nevertheless, this requires focus, a good understanding of the goals, and perpetual verification of the results.

A further obstacle to using LLMs in this stage of DEX modeling is the multiplicity of decision tables. A single model contains a separate table for each aggregate attribute. The model in Figure 1, for instance, contains 20 aggregate attributes and thus 20 distinct decision tables. LLM chatbots are not designed to manage this volume of structured data simultaneously. Even when tables are named explicitly, the underlying representations can change unpredictably [7]. Consequently, if we really want to manage decision rules ourselves, we cannot do this effectively in LLMs, but should use specialized MCDM software.

Table 3: Evaluation of the two best-ranked EV cars (ChatGPT)

Model	MG 4 Comfort 64		Citroën ë-C4	
Cost-Ownership	good	(low purchase-cost; running-cost low)	good	(very-low purchase-cost & running-cost)
Capability	good	(435 km WLTP, 140 kW DC, 363 L boot)	adequate–good	(416 km WLTP, 100 kW DC, 380 L boot)
Responsibility	good	(★ 5 Euro-NCAP, heat-pump optional, V2L 3.6 kW)	good	(★ 4 Euro-NCAP, heat-pump std., forthcoming V2L)
EV-SUITABILITY	GOOD		GOOD	
Rank	1		2	

7 Identification of alternatives

The goal of this stage is to identify decision alternatives and describe them in terms of values, assigned to input attributes of the model. This is another stage in which LLMs excel due to their ability to act like agents [18], autonomously collecting relevant data from external sources. In this case, the LLMs constructed lists of actual EV cars, consulting sources such as avto.net, citroen.si, reuters.com, and ev-database.org:

MG 4 Comfort 64 kWh – €31 490 → €24 290 after Eko-sklad
Citroën ë-C4 54 kWh – €31 900 → €24 700
Opel Corsa Electric 54 kWh – from €28 490 → €21 290
BYD Dolphin Surf Boost 43 kWh – €22 990 → €15 790*
Fiat 500e 42 kWh – promo €22 290 (price already after bonus)

It is important to understand that alternatives can be fully defined only after developing the model structure in stages A.1 to A.3. Namely, it is necessary to know the input attributes (i.e., terminal nodes of the hierarchy) and their value scales; only this allows the LLM to search for the right information and appropriately represent it for evaluating the alternatives. Without this background, LLMs still collect data from various sources, but the process is much more arbitrary and lacks control. We also observed that collected data is often old, even when updated sources are already available; it helps to explicitly instruct LLMs to look for recent data.

8 Evaluation of alternatives

In this stage, the five cars have been evaluated by the LLMs. All were evaluated as GOOD, except Fiat 500e, which was evaluated as ADEQUATE. Table 3 illustrates the output generated by ChatGPT, showing the evaluation results of the two best-ranked cars. In addition to the overall evaluation (EV-SUITABILITY), the output includes evaluations achieved at the three main branches of the model, together with a brief notification of reasons for such evaluations. This is an appropriate summary of evaluation results and corresponds to what decision analysts normally do in this stage.

9 Explaining the results of evaluation

In MCDM, it is essential to understand the evaluation method and explain the reasons underlying evaluation results. This is how ChatGPT explained the evaluation method:

How I applied the DEX model:

1. Leaf values were taken from the most recent Slovene price-lists or Europe-wide tech sheets (all ≤ €35 000 trims).
2. Running-costs use your 30 000 km/year mileage and the Q1-2025 average household tariff of €0.166 kWh-1.
3. Each quantitative item was mapped to the ordered value-scales we set earlier; the three rule-tables (“purchase-cost”, “running-cost”, “economy”) were then fired, and all higher nodes were combined with simple equal weights (0.4 Cost-Ownership | 0.4 Capability | 0.2 Responsibility).
4. Scores were coded excellent = 4, good = 3, adequate = 2, poor = 1; the root class was chosen by rounding the weighted mean to the nearest integer.

At first, everything seems to make sense. Items 1 and 2 correctly mention “leaf values”, in the sense of input attributes, explaining the sources and assessment principles. However, beyond that, careful reading reveals a number of disturbing facts. First, very few attributes are mentioned in the explanation, while the model discussed that far was much bigger. The question remains which attributes were actually considered in this evaluation: all or only those mentioned. Second, the explanation mentions “weights”, an important concept in quantitative MCDM methods, but of only secondary role in DEX [5]. Third, “scores” mentioned in explanation item 4 indicate that the LLM used a non-DEX value aggregation procedure, aggregating numeric scores instead of employing decision rules. This is different from that determined (and taken for granted by the user) in the previous stages.

Unfortunately, this often happens with LLMs, which tend to “forget” about the previous MCDM stages [7]. It is not uncommon that attributes, their value scales, decision rules and the evaluation method change from prompt to prompt. This severely undermines the trust in using LLMs and makes the whole process uneasy: rather than focusing on solving the decision problem, the user is forced to meticulously check each and every step. Also, it is not uncommon to discover logical errors or even basic computational errors (often referred to as “hallucinations”

[1]). In one of our sessions with ChatGPT, it displayed the evaluation formula

$$(0.2 \times 3) + (0.25 \times 4) + (0.15 \times 4) + (0.2 \times 3) + (0.15 \times 2) + (0.05 \times 2) = 3.15$$

which looked convincing, but gave a hard-to-notice wrong result; the correct result is 3.2.

10 Analysis of alternatives

The last stage of the MCDM process is the analysis of alternatives, which is aimed at exploring the decision space using methods such as what-if and sensitivity analysis. In principle, LLMs are capable of performing such analyses, giving appropriate answers and explanations to questions such as:

- *Carry out sensitivity analysis for Citroën ë-C3 and MG4 depending on buying price and operating costs.*
- *What would have to change for Fiat 500e 42 to become a good EV vehicle?*

For example, ChatGPT answers the second question by investigating several possible changes of input attribute values that improve the overall evaluation of Fiat 500e 42, summarizing the findings as follows:

Summary: what must change

1. **Stretch the body** (long-wheel-base or micro-estate) so boot ≥ 330 L and rear knee-room ≥ 800 mm.
2. **Approve a light-duty tow rating** and 50 kg roof rails.
3. **Raise DC fast-charge peak to 100 kW** with minor thermal upgrades.

None of those touches the battery size, motor, or purchase price, so **Cost-Ownership stays attractive**. Together they remove every "poor" leaf and elevate both *Practicality* and *Charging* just one step—exactly enough to push the Fiat 500e into the model's **GOOD** band.

A non-DEX evaluation mechanism was used again, but results are otherwise correct and informative. Nevertheless, the issues of using inappropriate model components and making logical and computational errors have been often observed in this stage as well.

11 A hybrid LLM-DEX workflow

Given the demonstrated strength of LLMs in the initial, conceptual stages of model development (A.1 – A.3), and their limitations in subsequent, more formal stages, we propose a hybrid workflow for DEX modelling:

1. *Model Structuring in LLM*: Develop the initial model structure, including the hierarchy of attributes and their value scales, using an LLM.
2. *Model Export*: Export this structural definition to a file.
3. *Model Completion in DEXiWin*: Import the file into the specialized software DEXiWin to carry out the remaining stages, notably the definition of decision rules (A.4) and the evaluation and analysis of alternatives (B.5 – B.8).

Implementing this workflow requires a reliable method for transferring the model from the LLM to DEXiWin. DEXiWin stores models in ".dxi" files [6], which are text files using a specific XML (eXtensible Markup Language [9]) schema to represent the model's attributes, decision tables, and alternatives.

While LLMs possess inherent knowledge of DEX methodology, we found they lack explicit knowledge of the .dxi file schema. Direct prompts, such as *Save the model to a .dxi file*, result in generically structured XML that is not readable by DEXiWin. Furthermore, providing a valid .dxi file as an example for the LLM to emulate also proved ineffective.

Table 4: Document Type Definition for exporting DEX models

```
<!ELEMENT DEXiFromLLM (criteria, alternatives)>

<!-- Criteria Tree -->
<!ELEMENT criteria (criterion+)>
<!ELEMENT criterion (description,
  (criterion | scale)*)>
<!ATTLIST criterion name CDATA #REQUIRED>

<!ELEMENT description (#PCDATA)>

<!-- Scale and qualitative values -->
<!ELEMENT scale (value+)>
<!ELEMENT value (#PCDATA)>
<!ATTLIST value description CDATA #IMPLIED>

<!-- Alternatives -->
<!ELEMENT alternatives (alternative+)>
<!ELEMENT alternative (data+)>
<!ATTLIST alternative name CDATA #REQUIRED>

<!-- Leaf criteria values in alternatives -->
<!ELEMENT data (#PCDATA)>
<!ATTLIST data criterion CDATA #REQUIRED>
<!ATTLIST data numeric CDATA #IMPLIED>
```

The successful approach utilized a Document Type Definition (DTD) file as a formal schema. A DTD precisely defines the structure and legal elements of an XML document. By providing the LLM with the appropriate DTD (Table 4), it generates an XML file with a well-defined and predictable structure. Although this XML is still not directly readable by DEXiWin, it can be reliably converted into a compatible ".dxi" format. Currently, this conversion is performed by a script, but this functionality is planned for integration into future releases of DEXiWin.

12 Discussion

LLMs are developing rapidly and becoming increasingly capable. They may evolve under the hood, so that even the same version can behave differently depending on recent updates or user-specific factors. This makes them challenging for conducting rigorous scientific research. They come without user manuals, requiring their users to explore their capabilities on their own.

This study is an experimental attempt to understanding the capabilities of the current (2025) mainstream LLMs for supporting the MCDM process, with special emphasis on the DEX method. On this basis,

Table 5: Recommendations for using LLMs in MCDM/DEX

Task	Recommendations
Problem Scoping and Ideation	**Highly recommended**. Use LLMs for collecting and brainstorming criteria, generating lists of alternatives, and exploring value scales.
Model Structuring	**Recommended with caution.** Use LLMs to propose initial hierarchical structures and define rough attributes. Requires verification and justification.
Preference (Decision Rule) Elicitation	**Not recommended.** Use only for initial brainstorming, but consider your own preferences. Use specialized software.
Defining Alternatives	**Recommended**. Use LLMs as agents once MCDM model has been defined.
Model Execution and Analysis	**Not recommended**. LLMs are unreliable for rigorous application of evaluation and analysis methods. Use specialized software.
Explanation and Reporting	**Recommended for drafting.** Use LLMs to help draft explanations of the methodology, summarize results in text, or generate reports.

we could not formulate firm conclusions, but were still able to make observations and formulate recommendations that might help MCDM practitioners.

The single most important contribution of LLMs to MCDM is their ability to formulate a well-structured list of relevant criteria in the first stage (A.1). Nothing nearly as good was available so far for that difficult stage. Now, LLMs can substantially boost the process and save a lot of effort and time. The second important contribution is the capability of LLMs to act as agents and collect data about alternatives (B.5) from various external resources.

Considering individual MCDM stages, LLMs performance is quite impressive. They are capable of evaluating and analyzing alternatives, without much instruction. Furthermore, if asked, they can explain the used methods and obtained results quite well. In some cases, however, a seemingly convincing explanation may fall apart, revealing logical, methodological and computational errors.

Considering the MCDM process as a whole, the performance of LLMs is not as favorable. In subsequent MCDM stages, LLMs tend to “change their mind” without notice, modifying the already established model components: attributes, value scales, decision rules and evaluation method. Consequently, this requires a lot of attention from the user’s side, who has to check the outputs and perpetually remind the LLMs to remain consistent. This distracts the process and often carries the user away of the main decision-making task. Also, we should warn that in the preference modelling stage (A.4), LLMs suggest generalized decision preferences that might substantially differ from the user’s subjective preferences, which need to be enforced explicitly.

This study evaluated LLMs on their ability to rigorously apply a specific MCDM method, DEX, to a complex decision problem. Our results demonstrate that this remains a significant challenge for current LLMs. However, in cases when we are less interested in the methodology and just want approximate answers to common decision problems (such as when buying a new mobile phone), LLMs turn out to be much better companions. In such scenarios, a few simple prompts are often sufficient to help users formulate their requirements, define criteria, and receive approximate, but useful, recommendations.

In summary, LLMs can substantially contribute to the definition of attributes and alternatives, but are unsuitable for carrying out the whole MCDM process due to possible inconsistent and erroneous executions of the MCDM method. Our findings suggest a pragmatic, collaborative approach where LLMs act as powerful assistants to the human decision modeler. The guidance is summarized in Table 5: use LLMs for divergent thinking and ideation in the early, creative phases of model development, but rely on established MCDM software, such as DEXiWin, and human expertise for convergent thinking and the rigorous, methodological execution of the decision-making process. Nevertheless, LLMs evolve fast and we may expect substantial improvements in the future.

For further research, we propose more detailed studies on the user experience with LLMs. Key questions include how users perceive the process, which functionalities they find most useful, what aspects distract them, their ability to detect and correct errors, and the level of methodological rigor they expect. This is an ideal task for the field of Behavioral Operations Research [10]. From the MCDM perspective, it would be valuable to compare LLM support for different methods. For instance, hierarchical methods like AHP and MACBETH share the initial problem structuring stages (A.1 and A.2) with DEX and may receive support of similar quality. As their subsequent stages diverge, a comparative analysis could identify differences in the quality of LLM support, assessing accuracy, logical consistency, and usability.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency for the programme Knowledge Technologies (research core funding No. P2-0103 and P5-0018).

References

[1] Bana e Costa, C., De Corte, J.-M., Vansnick, J.-C. (2003): MACBETH (Overview of MACBETH multicriteria decision analysis approach). *International Journal of Information Technology &*

Decision Making 11:02, 359–387. https://doi.org/10.1142/S0219622012400068.

[2] Banerjee, S., Agarwal, A., Singla, S. (2024). *LLMs will always hallucinate, and we need to live with this*. https://doi.org/10.48550/arXiv.2409.05746.

[3] Bansal, A. (2024): Comprehensive Study on LLM Agents. *International Journal of All Research Education & Scientific Methods* 12(8):2455-6211.

[4] Bohanec, M., Rajkovič, V., Bratko, I., Zupan, B., Žnidaršič, M. (2013): DEX methodology: Three decades of qualitative multi-attribute modelling. *Informatica* 37, 49–54.

[5] Bohanec, M. (2022): DEX (Decision EXpert): A qualitative hierarchical multi-criteria method. In: *Multiple Criteria Decision Making* (ed. Kulkarni, A.J.), Studies in Systems, Decision and Control 407, Singapore: Springer, https://doi.org/10.1007/978-981-16-7414-3_3.

[6] Bohanec, M. (2024): *DEXiWin: DEX Decision Modeling Software, User's Manual, Version 1.2*. Ljubljana: Institut Jožef Stefan, Delovno poročilo IJS DP-14747. https://dex.ijs.si/dexisuite/dexiwin.html.

[7] Bohanec, M. (2025): Using large language models for creating, modifying and interpreting decision tables in multi-criteria method DEX. *CECIIS 2025*, 36th Central European Conference on Information and Intelligent Systems, Varaždin, Croatia, 651–658.

[8] Bohanec, M., Rajkovič, U., Rajkovič, V (2025): Utilizing large language models for supporting multi-criteria decision modelling method DEX. *IS 2025*, Proc. of the 28th International Multiconference Information Society, Volume A, Ljubljana, Slovenia. 19–22.

[9] Bray, T., Paoli, J., Sperberg, C.M., Maler, E., Yergeau, F. (2008): *Extensible Markup Language (XML) 1.0* (Fifth Edition). https://www.w3.org/TR/REC-xml/.

[10] Hämäläinen, R.P., Luoma, J., Saarinen, E. (Eds.). (2016): *Behavioral Operational Research: A Capabilities Approach*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-25405-6.

[11] Ishizaka, A., Nemery, P. (2013): *Multi-criteria decision analysis: Methods and software*. Chichester: Wiley. ISBN: 978-1-119-97407-9.

[12] Kamath, U., Keenan, K., Somers, G., Sorenson, S. (2024): *Large Language Models: A Deep Dive: Bridging Theory and Practice*. Springer, 506p, ISBN-13 978-3031656460.

[13] Kulkarni, A.J. (Ed.) (2022): *Multiple Criteria Decision Making*. Studies in Systems, Decision and Control 407, Singapore: Springer, https://doi.org/10.1007/978-981-16-7414-3.

[14] Radovanović, S., Delibašić, B., Vukanović, S. (2024): Combining LLM and DIDEX method to predict Internal Migrations in Serbia. *Proc. 24th International Conference on Group Decision and Negotiation & 10th International Conference on Decision Support System Technology* (GDN ICDSST 2024), Vol. 1: Technology as a support tool (Eds. S.P. Duarte, P. Zaraté, A. Lobo, B. Delibašić, T. Wachowicz, M.C. Ferreira), University of Porto.

[15] Saaty, T.L., Vargas, L.G. (2012): *Models, Methods, Concepts & Applications of the Analytic Hierarchy Process*. Springer, US, New York, https://doi.org/10.1007/978-1-4614-3597-6.

[16] Svoboda, I., Lande, D. (2024): Enhancing multi-criteria decision analysis with AI: Integrating analytic hierarchy process and GPT-4 for automated decision support. https://arxiv.org/abs/2402.07404.

[17] Wang, H., Zhang, F., Mu, C. (2025). One for all: A general framework of LLMs-based multi-criteria decision making on human expert level. arXiv e-prints, 2502.15778. https://arxiv.org/abs/2502.15778.

[18] Wang, L., Ma, C., Feng, X. et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 186345. https://doi.org/10.1007/s11704-024-40231-1.

[19] Yuan, W., Liu, P., Gallé, M (2024): LLMCRIT: Teaching large language models to use criteria. *Findings of the Association for Computational Linguistics*: ACL 2024, 7929–7960, https://doi.org/10.18653/v1/2024.findings-acl.472.