Paradigm

A Seven-Criteria Framework to Evaluate Initiatives for Generative AI-Based Solutions

Marius SAVA*
National University of Science and Technology Politehnica, Bucharest, Romania
**Corresponding author, marius.sava@stud.acs.upb.ro*

Gheorghe MILITARU
National University of Science and Technology Politehnica, Bucharest, Romania
gheorghe.militaru@upb.ro

Abstract. *The development and potential of large language models created incentives for organizations to adopt and integrate this technology in their internal processes. When the technological layer of organizations is in an acceleration phase there is a strong need to manage this change. Currently, generative artificial intelligence brings a vast area of opportunities, but this can create an "analysis paralysis" when deciding what initiatives to pursue. Different frameworks exist but they are targeting software solutions or specific domains. This paper follows an exploratory, framework-development qualitative design. An evaluation framework to assess initiatives based on generative artificial intelligence technology is developed by reusing existing decision models and enhancing them with criteria based on a simulated focus group using large language model personas. The interim findings were analysed using directed qualitative content analysis (deductive coding to a baseline codebook derived from literature), complemented by qualitative methods (free listing of evaluation criteria, anchored rating-scale elicitation, priority ranking, mapping to baseline criteria). A derived seven-criteria framework with definitions and performance scales resulted. This allows organizations to maximize chances of choosing the highest-reward initiatives and invest resources efficiently when needed to select, implement and integrate generative artificial intelligence solutions. The intention of this paper is to propose a list of seven evaluation dimensions — domain effectiveness and impact, regulatory compliance and ethical risk, user adoption, business value, technical feasibility, data availability and quality, and competitive differentiation that can be used to characterize applications based on this technology. [ms1][ms2]*

Keywords: generative artificial intelligence, portfolio decision making, initiative selection, evaluation framework, large language models, organizational change, resource allocation.

Introduction

Since the emergence of Large Language Models (LLMs), a core building block of generative AI, organizations across many private and public sectors have experienced a new phase of evolution and transformation. From generic supporting services that assist staff in many activities such as translation, summarization, and classification to highly customized assistants for individuals that can recommend products tailored to specific needs, schedule calendar meetings or suggest courses of action based on organizational data, this technology is increasingly common in organizational activities as well as in personal life.

Large language models were rapidly adopted as the main generative AI technology in organizations of different sizes. Recent developments, such as open-weight models, better integration of agentic AI and the increasing ease of developing technological solutions, create a fostering environment able to generate added value; however, they also create an unfavourable context for organizations, given the large number of initiatives that can be pursued leading to

DOI: 10.2478/picbe-2026-0121

opportunity costs. Choosing the most optimal initiatives is a challenging task, and there are no complete and safe tools that can be used to predict outcomes. A general framework to choose wisely in a timely manner, from a large pool of initiatives based on generative artificial intelligence is necessary. Well established frameworks and de facto standards for industry like the health technology assessment (HTA) Core Model (Kristensen et al., 2017) or the general model for information-systems, the IS-Success model (DeLone & McLean, 2003) are useful models to decide whether a solution is worth pursuing or not. However, they do not meet the need for generic, fast and reusable solutions given the rapid change in today's generative AI field. In the next sections, the current identified models are compared together with their advantages and disadvantages. Based on this analysis, the paper creates a synthesis in the form of a new framework, which is refined through a simulated AI focus-group.

Finally, the article proposes a set of seven evaluation dimensions that can be used to characterize applications based on this new technology.

Literature review

Contemporary work on generative AI largely falls under the "*acting humanly*" perspective on the concept of intelligence, a viewpoint in which systems are evaluated primarily by how closely their behaviour resembles that of humans. Generative AI are systems which can produce new content (e.g. text, images, videos) based on previous large content they were exposed to, through a process called training.

A literature search was performed using the Web of Science Core Collection and Google Scholar. The search strategy encompassed two main phases. The first one focused on identifying relevant articles for generative AI evaluation frameworks. Multiple keyword phrases were used such as "*generative AI evaluation, large language models evaluation, information system evaluation, decision making framework*", each combined with domain specific keywords including "*marketing*", "*sustainability*" and "*health*". Following these queries a selection of 120 potential articles was identified via title screening. A further screening of abstracts reduced the number to 35 papers. The second phase consisted of using a citation tracking process, resulting in the identification of two foundational frameworks relevant for this paper.

The health technology assessment (HTA) Core Model (Kristensen et al., 2017) and the general model for information-systems, the IS-Success model (DeLone & McLean, 2003), were identified as solid model options. The HTA Core Model is useful when a comprehensive evaluation needs to take place covering a multitude of criteria such as safety, costs, organizational aspects, patient and social aspects, legal aspects and ethics. While this approach provides benefits for generative AI initiatives, it is slow and heavy to be applied and because it was designed for the healthcare sector it needs to be heavy adapted. The IS Success Model is useful because it links all important factors for an information system to be successful such as quality to use, business value and user benefits. For generative AI initiatives, it will explain why some solutions will fail or succeed in practice, but it does not explicitly cover aspects such as the level of compliance or ethics.

Methodology

This paper uses an exploratory qualitative framework-development design, the main goal being to identify the most relevant criteria which can form the basis for an actionable evaluation framework that can be used by organizations as a filter, given the immense number of generative AI initiatives which can be pursued. The qualitative design was chosen because a directed qualitative content analysis approach allows reuse of existing work while remaining flexible enough to adapt it for

new uses, such as evaluating generative AI initiatives. This framework proposes a set of equally important criteria. All measures indicating relative importance are interpreted in a descriptive way, therefore a multi-criteria analysis is not within scope in this paper.

The methodology to develop the framework followed a six-step approach with the phases being presented in Figure 1. Initially, a literature review was performed in order to identify relevant studies which can be analysed. After these studies were analysed and synthesised, an interim set of criteria was prepared. The next phase was to validate the set of criteria through a simulated focus group. This step collected a comprehensive set of data from multiple perspectives. Finally, multiple qualitative techniques were applied in order to analyse the focus group data and a final, seven-criteria framework including definitions and performance scales was developed.

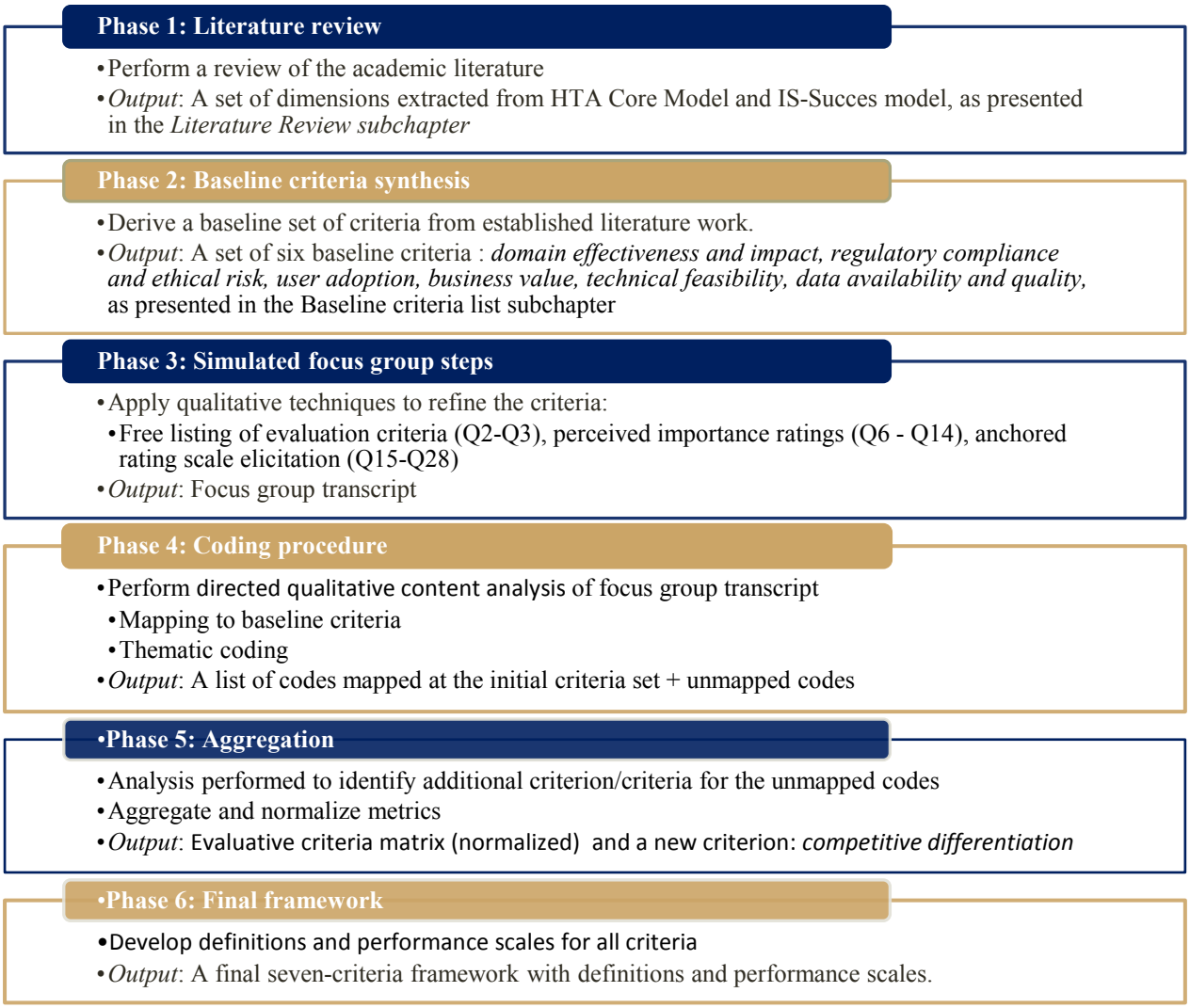


Figure 1. Research methodology used in this paper

Source: Authors' own research results.

The refinement of the criteria is completed using a simulated focus group approach. A simulated focus-group was chosen due to speed, high levels of accessibility to experts, and lower costs when comparing to a human focus group. Because outputs from large language models have

known issues such as bias or hallucination, mitigation measures were taken such as using the transcript only for extracting candidate criteria, manually checking transcript for bias or hallucination, and providing an extensive interview guide.

There are three research objectives for the focus group: identifying the set of criteria, a description of them and a scale together with an elicitation of a criteria profile based on descriptive metrics such as frequency. The focus group methodology is applied as described by (Krueger & Casey, 2000) with the consideration that the focus group is a simulated one. In any focus group lifecycle there are three main phases: one being the preparation part, where the interview guide was developed, followed by the performing of the discussion, in this case, running the simulation, and closing with the analysis part, which involves the application of several well-known techniques in order to extract value from the obtained data.

Baseline criteria list

Six baseline evaluation criteria are derived by synthesizing two mature lines of work: the health technology assessment (HTA) Core Model (Kristensen et al., 2017) and a general model for information-systems, the IS-Success model (DeLone & McLean, 2003).

The HTA Core Model specifies a multidimensional structure for medical technology evaluation appraisal with distinct constructs for technical characteristics, safety and ethical concerns, costs, including patient and organizational aspects. The (DeLone & McLean, 2003) model highlights as core constructs the quality of a system, implicitly its information and service one, the act of using the actual system together with user satisfaction, and quantifying the outcomes as net benefits.

Regulatory compliance and ethical risk are foundational criteria in HTA Core Model. Technical feasibility encompasses the HTA domain of technical characteristics, function or safety with the overall approach of "system quality" in the IS-success model, which covers reliability, performance and usability. *Data availability and quality* is a mandatory layer to be evaluated, being of utmost relevance for AI-specific solutions, like information quality criteria in the IS-success list (accuracy, completeness) and of the evidence base required for HTA Core Model. *Business value* aligns closely with the "costs and economic evaluation" domain in the HTA Core Model, and with the net benefits construct in the IS-success model. *Domain effectiveness and impact* generalizes HTA's clinical effectiveness but also corresponds to the "organizational impact" construct in the IS-success model. Finally, *user adoption* is grounded in technology-acceptance of IS-success model. Thus, the six baseline criteria combine high-order criteria of HTA Core Model and IS-success constructs.

Conducting the focus group

The paper employs a prompt-driven AI simulated of a focus group (Zhang et al., 2024) to establish criteria for assessing generative AI product initiatives, using AI personas of Romanian nationality with diverse professional backgrounds such as healthcare, marketing, and sustainability. The simulation is executed via the Simulation Focus Group platform (Focus Groups - AI, n.d.). The parameters used as input for the platform are the ones presented in Table 1. Moderation prompts were generated automatically by the platform. One simulation was performed as the subsequent analysis is completely manual.

Table 1. Configuration parameters for the Focus Group AI platform

Parameter name	Value
Version	2026 FocusGroups-AI
Type of setup	Manual setup
Purpose	to identify the most relevant criteria which can form the basis for an actionable evaluation framework that can be used by organizations as a filter, given the immense number of initiatives which can be pursued across diverse sectors such as healthcare, marketing, and sustainability
Topic	filtering generative AI initiatives
Test messages	verbatim questions from the Appendix
Personas	Participant 1 is a 38-year-old male Romanian who works as a software delivery lead for a global IT company. Participant 2 is a 44-year-old Romanian female with postgraduate education, employed as a data protection officer. Participant 3 is a 33-year-old Romanian female with a bachelor's degree, leading market strategy for a telecommunications company. Participant 4 is a 41-year-old male Romanian physician who practices internal medicine and has a postgraduate degree. Participant 5 is a 36-year-old Romanian woman with a bachelor's degree who works as an expert in the sustainability domain, specialising in renewable energy projects.

The design comprises four iterative steps as shown in Figure 1, phase 3. First, a free-listing phase (Weller & Romney, 1988) captures the space of evaluation criteria from each persona. Second, a baseline mapping activity organizes these criteria into a predefined authoritative list, using coding technique (Bingham, 2023). Third, an anchored rating-scale elicitation technique is used to ask participants to rate the importance of each criterion on a defined scale (Likert, 1932). Finally, all criteria are organized based on aggregated judgements and descriptive anchors are created (Smith & Kendall, 1963). These techniques are covered by a sequence of 28 scripted questions in current focus group interview guide.

Data analysis approach

The two sources for deriving and structuring evaluation criteria are the synthetic data obtained from the AI-simulated focus group and the data obtained through author research for creating the baseline list of criteria, as shown in phase 1, Figure 1. Six synthesized criteria were used as the base for the directed qualitative content analysis.

Using thematic coding (Braun & Clarke, 2006), all sentences that expressed evaluative remarks about generative AI initiatives were selected and coded manually by the author, at the level of clauses. If a sentence contained multiple remarks, it was split into separate parts. Evaluative remarks regarding generative-AI initiatives were assigned descriptive codes and correlated with the six baseline criteria, with the rule that each clause could receive only one code. When a clause could not be mapped, this resulted in the creation of a new code. At the end of the coding process one new criterion was derived, namely, competitive differentiation.

For each criterion, the frequency of mentions (how many clauses were mapped to a criterion), average importance ratings as indicated by personas, and instances where a criterion was

selected as a top priority by each persona were calculated in the matrix described in Table 2. This matrix created the premises to create a profile that describes each criterion.

Selecting group

A major success factor for achieving a comprehensive analysis is the presence of individuals possessing the appropriate domain knowledge for the specific situation. For this paper, the AI participants profiles were defined in such a way as to target general competences like product development and compliance, but also domain specific knowledge.

Identifying participants involves two steps. First, domains are mapped to personas: for each of the three domains (healthcare, market-intelligence, sustainability), a persona is defined so that domain expertise is well represented. Then, role coverage is checked to ensure that the key roles (software delivery lead, data protection officer, market lead and subject-matter expert) are present, so that no core perspective on the decision problem is missing. Five personas were provided in the configuration setup utilised in this focus group as defined in Table 1. All persona weights were considered equal given the simulated environment, where it is assumed that the personas, based on their description, will have similar strength of arguments during discussion. Together, these personas span the necessary perspectives for this paper objectives.

Developing interview guide

The interview guide is presented in the Appendix and was created based on the focus group objectives. Developing the interview guide is part of the overall planning phase and requires appropriate tools, a clear conceptual framework and methodological rigour. The starting point is the research question: the guide must elicit what matters regarding the criteria, how these aspects can be observed and measured. Five thematic blocks implemented as a set of 28 questions were structured according to (Krueger, 1998), each with optional follow-up probes where clarification or elaboration was needed. The organizing schema used for categorizing the questions follows.

Q1 records participant characterisation, providing contextual information on the stakeholder role. Using free listing approaches to map the domain and extract insights from personas, Q2–Q3 create an open-ended free listing of evaluation criteria (Weller & Romney, 1988). These are further integrated through a mapping procedure against a predetermined, authoritative list. Q4 and Q5 ask personas to examine their elicited criteria in order to facilitate the evaluation of the candidate criteria list. A 5-point Likert-type scale (from 1 = "*Not at all important*" to 5 = "*Extremely important*") is used in Q6–Q12 to assess perceived importance ratings across categories, while Q13–Q14 ask personas to rank priorities (Parnell & Trainor, 2009). Lastly, the development of anchored descriptions for scale extremes is facilitated by Q15–Q28 (Martin-Raugh et al., 2016).

Document and evaluate

The simulated focus group is executed on a web-based platform for scripted dialogues, which allows the researcher to define personas and their profiles, specify the moderator agent together with the interview guide and generate a full transcript of the conversation.

All simulated sessions are automatically registered by the platform, and the resulting transcripts are exported for further processing. Firstly, a thematic coding is performed. Raw answers are interpreted into codes which are further mapped to baseline criteria or to newly criteria. An initial description was created for each dimension by abstracting the relevant IS-Success and HTA Core Model criteria.

These provisional definitions were then empirically refined using the focus-group data: the transcripts were thematically coded, and all codes were mapped onto the candidate dimensions through an iterative process. There were codes that could be mapped to competitive advantages area, a new dimension which is different from the baseline. Based on this, a seventh criterion, competitive differentiation, was introduced. Each criterion is described by a refined definition accompanied by mentions, inclusion rates, importance ratings, swing priorities and performance scales summarized in Table 2.

Results and discussions

The key findings of the focus-group simulation are summarised in this section. The final evaluation criteria are described and compared against each other (mentions, inclusion rate, importance ratings and swing priorities). The results are then interpreted qualitatively, as presented in the following sections.

Table 2. Evaluative criteria and their definition

Criteria name	Baseline /derived	Mentions	Definition
domain effectiveness and impact	baseline	37	The extent to which the solution produces measurable improvements (quality, safety, accuracy) compared to baseline.
regulatory compliance and ethical risk	baseline	28	Alignment of the implementation with legal requirements (privacy, safety).
user adoption	baseline	19	Likelihood that intended users will use the solution in their real workflows, given their incentives and skills.
business value	baseline	16	Expected economic value (revenue, savings, strategic positioning) is relative to the investment.
technical feasibility	baseline	12	Extent to which the solution can be built, integrated, and operated with current infrastructure and resources.
data availability and quality	baseline	3	Availability and fitness-for-purpose of data (coverage, labeling, bias, noise) needed to train the models.
competitive differentiation	derived	1	Degree to which the solution creates a sustainable edge (e.g. proprietary data, integrations) that makes it harder for competitors to develop.

Source: Adapted from (DeLone & McLean, 2003; Kristensen et al., 2017).

The starting point for the analysis was the six-criteria baseline list. After the focus group transcripts were thematically coded, each coded segment was either mapped to one of these baseline criteria or, in the absence of a suitable match, to a temporary category.

As presented in Table 2, six criteria were adapted based on IS-Success model and HTA Core Model, while one criterion—competitive differentiation—was derived from codes that referred to advantage and difficulty for competitors to replicate it. The table contains, for each criterion, a definition, the number of mentions (indicating how many distinct codes referred to that criterion), across the transcripts, which gives an indication of relevance.

Table 3. Evaluative criteria matrix (normalized)

Criterion	Mention	Inclusion rate (Q4)	Top 3 frequency (Q5)	Mean importance (Q6–Q12)	Swing priority frequency (Q13)	Swing priority frequency (Q14)
domain effectiveness and impact	0.32	0.14	0.13	0.15	0.80	0.20
regulatory compliance and ethical risk	0.24	0.23	0.33	0.17	0.20	0.80
user adoption	0.16	0.14	0.13	0.15	0.00	0.00
business value	0.14	0.18	0.13	0.14	0.00	0.00
technical feasibility	0.10	0.09	0.13	0.11	0.00	0.00
data availability and quality	0.03	0.23	0.13	0.17	0.00	0.00
competitive differentiation	0.01	0.00	0.00	0.11	0.00	0.00

Source: Author's own research contribution.

Based on these metrics, Table ***3*** summarizes the profiles of each criterion. *Domain effectiveness and impact* and *regulatory compliance and ethical risk* were continuously highlighted in free listings phase or swing priority questions. *User adoption* and *business value* are the next most important, a translation of the fact that each product needs to have a market. *Technical feasibility* and *data availability and quality* criteria are discussed as necessary enablers, indicating that they need to be taken into consideration. Although it is rarely discussed, *competitive differentiation* is kept as a separate criterion since it expresses a clear concern about long-term success.

The evaluative criteria matrix's purpose is to provide a structural insight for assessing framework coherence. The table is sorted by mentions, for convenience without using it as a core metric for ordering. Mentions show a weak correlation with the inclusion rate (Q4). The low inclusion rate (Q4) for the *domain effectiveness* criterion is likely an alignment bias effect of the AI training dataset and strong conformance rules. The AI platform is set to give more protective answers to questions so there is an inflation effect caused by these settings which were outside of the author's control. This is mostly visible during the Q4 question where the personas were required to choose and sacrificed "*business value*" of any kind for security. The overly prudent approach of the platform may also affect the other metrics, but the effect remains unquantified. The other outlier is data availability, which was mostly unmentioned, but when it was presented as a choice, it becomes evident that it is mostly a prerequisite without which no project could succeed as shown by the mean importance rating (Q6-Q12). That is, data existence is treated as a prerequisite from the personas' perspective.

Top three frequency (Q5) and swing priority frequency (Q13-Q14) suffer from the security inflation effect with overemphasis *on regulatory compliance and ethical risk* criterion. For the mean importance rating (Q6-Q12) the scores were equally distributed, and this effect could be explained by the perspectives of personas. In this metric, each criterion was rated as important, for example, *competitive differentiation* being scored higher by Participant 3 – the lead in market field.

The c*ompetitive differentiation* criterion has low mentions, but after the analysis of all the other criteria, it was concluded that it should be used as a strategic enabler because it is the single criterion which takes into consideration long-term thinking, creating a market advantage. While an

initiative can check all the other criteria, it may fail rapidly as a business because too many companies are doing similar things.

While multiple effects were involved in the final matrix values, the implications for the framework are that these scores shall not be used for a multi-criteria analysis. A separate analysis shall be conducted to determine grounded weights for the criteria.

The criteria and associated scales were developed with the principle of minimum conceptual overlap. The most difficult to separate cluster is *domain effectiveness and impact*, *business value* and *user adoption.* This potential conceptual adjacency was addressed by assuring that each criterion evaluates different scale dimensions. The *domain effectiveness and impact* criterion asks for a specific improvement in organizational metrics (e.g. quality or accuracy), while the *business value* is evaluated only financially. For example, an initiative may have high operational impact without immediate profit foreseen. It must be recognised that while improvements may generate greater financial value the two distinct dimensions are measured separately. In the same manner, the *user adoption* criterion, quantifying the actual integration into business workflows, is not a subset of generated business value. At the moment of evaluation of generative AI initiatives, this criterion captures a distinct, quantifiable aspect of the solution which is not directly monetizable.

The final step in the analysis is to transform these qualitative profiles into operational 0–5 performance scales for each criterion. Participants were asked to describe "*very poor*" and "*excellent*" situations for each criterion during the discussion. Using the same underlying logic across criteria, these narrative anchors were aggregated into descriptions for scores 0–5 where 0 refers to the absence of the criteria and score 5 represents an exemplary level for that criterion. This resulted in a set of scales that can be applied uniformly across various generative AI initiatives.

In this evaluation, the personas represent professionals with similar levels and the same nationality, specifically Romanian. It must be considered that Romania, being a member of the European Union, is heavily influenced by the GDPR, likely having an elevated effect on regulatory compliance. This indicates that it may be the case that other criteria will be emphasised more in international contexts where regulatory laws are less restrictive. Industry also plays a significant role when applying the framework. A simple mechanism could be used, where some criteria will become mandatory depending on context. In sectors heavily regulated (Finance, Health), criteria like *regulatory compliance and ethical risk* and *data quality and availability* will play a significant role as a hard constraint for any generative AI initiative, while less restrictive sectors will be influenced by criteria like *user adoption* and *competitive differentiation*.

Table 4. Evaluative criteria and their associated scale definition

Criterion	Scale definition
domain effectiveness and impact	0 – no improvement.
	1 – small improvement.
	2 – some measurable improvement in a few areas.
	3 – clear improvement in key metrics.
	4 – large improvements across most metrics.
	5 – a radical improvement
regulatory compliance and ethical risk	0 – multiple formal approvals required
	1 – one major approval
	2 – registration-based mechanism

Criterion	Scale definition
	3 – soft compliance
	4 - minimal approvals
	5 – no formal approvals required
user adoption	0 – no adoption
	1 – used rarely
	2 – some users use it occasionally
	3 – regular use by a large share of users
	4 – widely use
	5 – near-universal adoption
business value	0 – no credible revenue
	1 – less than €25k/yr
	2 – €25k–€100k/yr
	3 – €100k–€500k/yr
	4 – €500k–€2M/yr
	5 – more than €2M/yr
technical feasibility	0 – no minimum viable product (MVP)
	1 – an MVP can be created over 24 months
	2 – an MVP can be created in 12 – 24 months
	3 – an MVP can be created in 6 – 12 months
	4 – an MVP can be created under 6 months
	5 – an MVP can be created under 3 months
data availability and quality	0 – datasets scarce or inaccessible
	1 – fragmented data
	2 – data accessible, but heavy cleaning required
	3 – sufficient data
	4 – data that are organized
	5 – datasets abundant and ready to be shared
competitive differentiation	0 – open-source alternatives already exist with similar features
	1 – minor differences only
	2 – at least one feature is novel, but easy-to-copy
	3 – at least one feature is novel, but hard-to-copy
	4 – access to data replicating the work is very hard
	5 – distinct technical approach supported by hard access to data

Source: Author's own research contribution

Robustness and validation limitations of the framework must be considered when the framework is applied. Because the framework is derived using a small set of AI personas, it should be externally validated using a validation plan. For a grounded validation, human experts must be involved by conducting at least one human focus group. A further step for enhancing the strength of the framework is using the Delphi method. While it is a very complex method it creates the most

trustworthy results. The last factor which needs to be taken into consideration for the validation plan is the robustness of the framework. Multiple simulation setups must be performed using different persona sets or platform configurations monitoring the stability of the criteria over the iterative process.

Conclusion

In this article a criteria set, grounded both in established frameworks and in a simulated focus-group activity was developed, identifying seven criteria relevant to prioritise generative AI initiatives in any organization. For each criterion a profile with definitions and performance scales was created. The following criteria were domain effectiveness and impact, regulatory compliance and ethical risk, user adoption, business value, technical feasibility, data availability and quality, and competitive differentiation.

The framework offers a coherent perspective for organizations when considering generative AI initiatives given the fast pace at which technology is changing. The criteria can be used to articulate quickly where the opportunities are and which are the main risks. It is a fast and easy-to-use instrument for quick appraisal. The main limitation of this paper is the use of LLM-based personas in place of human participants which may introduce bias or hallucinations in spite of precautionary measures taken. For future work, the stability across multiple simulations of the simulated focus group may be assessed or complemented with human-based focus groups.

Acknowledgments

This work was supported by funds from the National University of Science and Technology Politehnica Bucharest, Romania.

References

Bingham, A. J. (2023). From data management to actionable findings: A five-phase process of qualitative data analysis. *International Journal of Qualitative Methods*, *22*, 16094069231183620. https://doi.org/10.1177/16094069231183620

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

DeLone, W. H., & McLean, E. R. (2003). The DeLone and McLean model of information systems success: A ten-year update. *Journal of Management Information Systems*, *19*(4), 9–30. https://doi.org/10.1080/07421222.2003.11045748

Focus Groups - AI. (n.d.). *AI-powered focus group simulator: Get consumer insights in minutes*. Retrieved 31 January 2026, from https://focusgroups-ai.com/

Kristensen, F. B., Lampe, K., Wild, C., Cerbo, M., Goettsch, W., & Becla, L. (2017). The HTA Core Model®—10 years of developing an international framework to share multidimensional value assessment. *Value in Health*, *20*(2), 244–250. https://doi.org/10.1016/j.jval.2016.12.010

Krueger, R. A. (1998). *Developing questions for focus groups*. SAGE Publications, Inc. https://doi.org/10.4135/9781483328126

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research*. SAGE.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 1–55.

Martin-Raugh, M., Tannenbaum, R. J., Tocci, C. M., & Reese, C. (2016). Behaviorally anchored rating scales: An application for evaluating teaching practice. *Teaching and Teacher Education*, *59*, 414–419. https://doi.org/10.1016/j.tate.2016.07.026

Parnell, G. S., & Trainor, T. E. (2009). Using the swing weight matrix to weight multiple objectives. *INCOSE International Symposium*, *19*(1), 283–298. https://doi.org/10.1002/j.2334-5837.2009.tb00949.x

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*(2), 149–155. https://doi.org/10.1037/h0047060

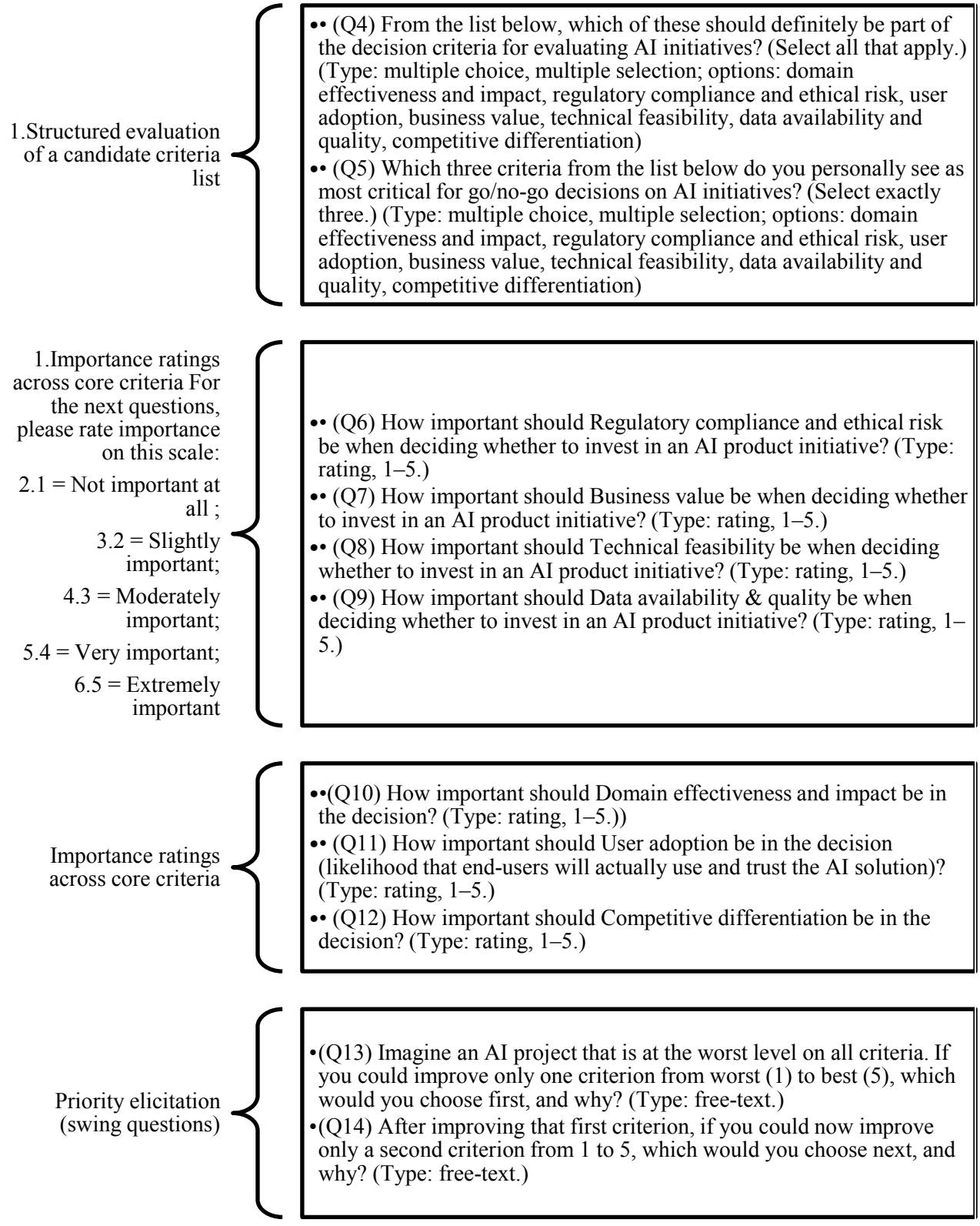
Weller, S. C., & Romney, A. K. (1988). Defining a domain and free listing. In *Systematic data collection* (pp. 10–20). SAGE Publications, Inc. https://doi.org/10.4135/9781412986069

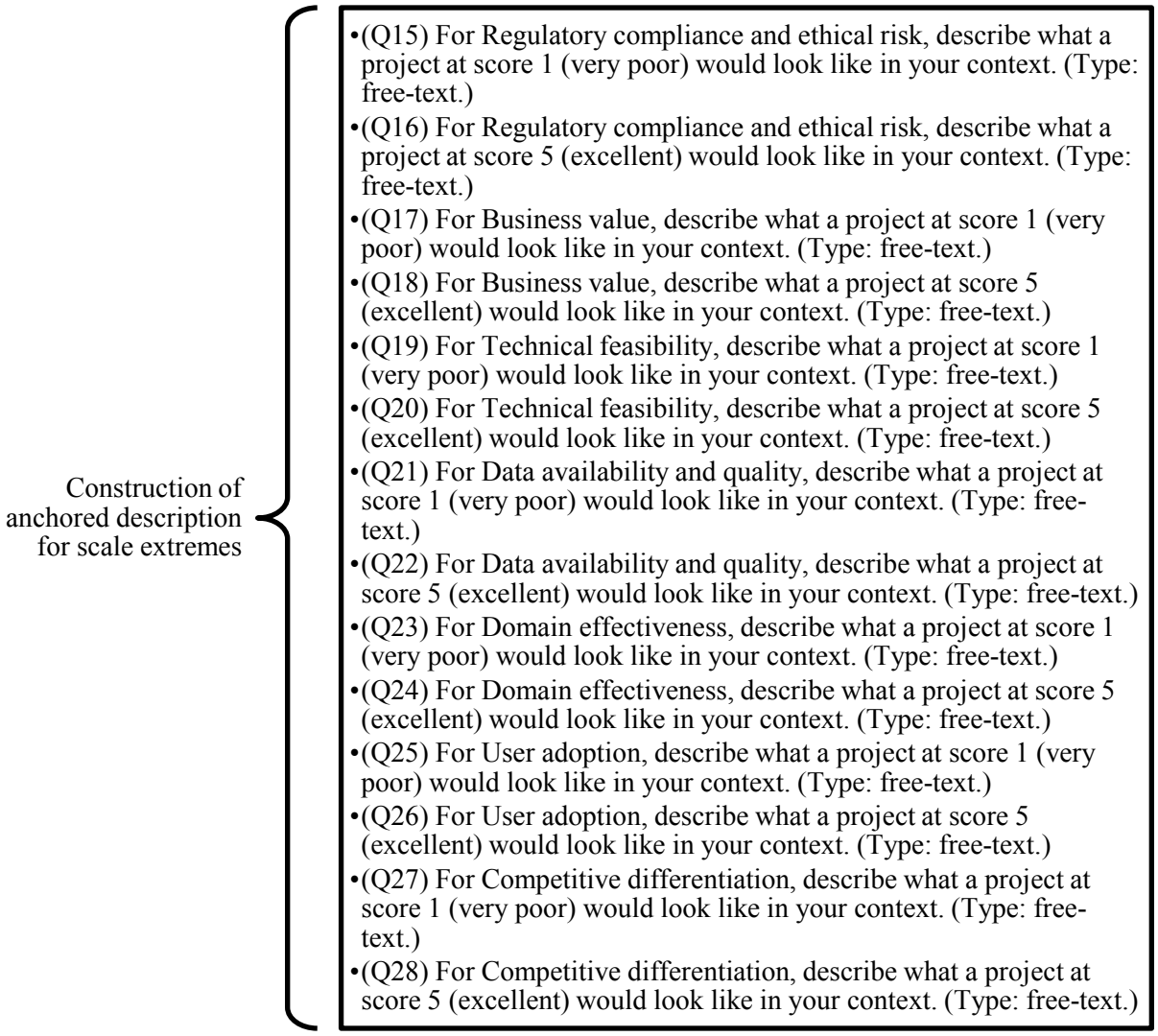
Zhang, T., Zhang, X., Cools, R., & Simeone, A. (2024). Focus agent: LLM-powered virtual focus group. *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents, IVA '24*. https://doi.org/10.1145/3652988.3673918

Appendix

Interview guidance questions (Source: Author's own research contribution)

Participant characterisation	•• (Q1) Which role best describes your perspective in this discussion? (Type: multiple choice; example options: Software delivery; Data protection officer; Market lead; Physician; Sustainability expert.)
Open-ended elicitation of criteria	•• (Q2) In your own words, what would count as a successful generative-AI product initiative in your organisation over the next 3–5 years? (Type: free-text.) •• (Q3) List up to five factors you consider essential when deciding whether to invest in a new AI-based product idea. (Type: free-text.)


1.Structured evaluation of a candidate criteria list
•• (Q4) From the list below, which of these should definitely be part of the decision criteria for evaluating AI initiatives? (Select all that apply.) (Type: multiple choice, multiple selection; options: domain effectiveness and impact, regulatory compliance and ethical risk, user adoption, business value, technical feasibility, data availability and quality, competitive differentiation)
•• (Q5) Which three criteria from the list below do you personally see as most critical for go/no-go decisions on AI initiatives? (Select exactly three.) (Type: multiple choice, multiple selection; options: domain effectiveness and impact, regulatory compliance and ethical risk, user adoption, business value, technical feasibility, data availability and quality, competitive differentiation)
1.Importance ratings across core criteria For the next questions, please rate importance on this scale:
2.1 = Not important at all ;
3.2 = Slightly important;
4.3 = Moderately important;
5.4 = Very important;
6.5 = Extremely important
•• (Q6) How important should Regulatory compliance and ethical risk be when deciding whether to invest in an AI product initiative? (Type: rating, 1–5.)
•• (Q7) How important should Business value be when deciding whether to invest in an AI product initiative? (Type: rating, 1–5.)
•• (Q8) How important should Technical feasibility be when deciding whether to invest in an AI product initiative? (Type: rating, 1–5.)
•• (Q9) How important should Data availability & quality be when deciding whether to invest in an AI product initiative? (Type: rating, 1–5.)
Importance ratings across core criteria
••(Q10) How important should Domain effectiveness and impact be in the decision? (Type: rating, 1–5.))
•• (Q11) How important should User adoption be in the decision (likelihood that end-users will actually use and trust the AI solution)? (Type: rating, 1–5.)
•• (Q12) How important should Competitive differentiation be in the decision? (Type: rating, 1–5.)
Priority elicitation (swing questions)
•(Q13) Imagine an AI project that is at the worst level on all criteria. If you could improve only one criterion from worst (1) to best (5), which would you choose first, and why? (Type: free-text.)
•(Q14) After improving that first criterion, if you could now improve only a second criterion from 1 to 5, which would you choose next, and why? (Type: free-text.)


Construction of anchored description for scale extremes
•(Q15) For Regulatory compliance and ethical risk, describe what a project at score 1 (very poor) would look like in your context. (Type: free-text.)
•(Q16) For Regulatory compliance and ethical risk, describe what a project at score 5 (excellent) would look like in your context. (Type: free-text.)
•(Q17) For Business value, describe what a project at score 1 (very poor) would look like in your context. (Type: free-text.)
•(Q18) For Business value, describe what a project at score 5 (excellent) would look like in your context. (Type: free-text.)
•(Q19) For Technical feasibility, describe what a project at score 1 (very poor) would look like in your context. (Type: free-text.)
•(Q20) For Technical feasibility, describe what a project at score 5 (excellent) would look like in your context. (Type: free-text.)
•(Q21) For Data availability and quality, describe what a project at score 1 (very poor) would look like in your context. (Type: free-text.)
•(Q22) For Data availability and quality, describe what a project at score 5 (excellent) would look like in your context. (Type: free-text.)
•(Q23) For Domain effectiveness, describe what a project at score 1 (very poor) would look like in your context. (Type: free-text.)
•(Q24) For Domain effectiveness, describe what a project at score 5 (excellent) would look like in your context. (Type: free-text.)
•(Q25) For User adoption, describe what a project at score 1 (very poor) would look like in your context. (Type: free-text.)
•(Q26) For User adoption, describe what a project at score 5 (excellent) would look like in your context. (Type: free-text.)
•(Q27) For Competitive differentiation, describe what a project at score 1 (very poor) would look like in your context. (Type: free-text.)
•(Q28) For Competitive differentiation, describe what a project at score 5 (excellent) would look like in your context. (Type: free-text.)