APPLICATION OF LARGE LANGUAGE MODELS IN DECISION SUPPORT SYSTEMS. PART II: Measurement and Assessment of Decision-Maker's Satisfaction

A. A. Kulinich

Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia

✉ alexkul@rambler.ru

Abstract. This paper is a continuation of the multi-part study devoted to the application of large language models (LLMs) in decision support systems (DSSs). Part II considers the issues of measuring the quality of a hybrid DSS that includes a human decision-maker (DM) and an LLM as an assistant. Many criteria for assessing the quality of a hybrid DSS and an expert assessment procedure are proposed. A key feature of the expert assessment of the system's quality is a structural model of a reasonable expert; this model is used to carry out a quasi-experiment for assessing the hybrid system. The assessment is performed within Five Whys, a well-known problem-solving methodology. An example of the operation and assessment of a hybrid DSS is provided, and the experiment results are analyzed. The new method for measuring quality and assessing the DM's satisfaction, based on the structural model of a reasonable expert and a quasi-experiment, can be useful at the development stage of a DSS that incorporates an LLM. This method yields high-quality assessments of hybrid DSSs while reducing the time and cost of assessment without engaging expert groups.

Keywords: decision support, large language model (LLM), hybrid system, assessment criteria, structural model of a reasonable expert.

INTRODUCTION

Part I of the study has addressed the issues of using large language models (LLMs) in decision support systems (DSSs) as an explanatory assistant in the processes of analyzing the decision situation, generating alternative solutions and making an appropriate choice, justifying the solution, and implementing it [1]. Part II is focused on assessing the interaction between a human and an LLM (artificial intelligence, AI). In this case, the matter concerns human–machine interaction and the assessment of its quality.

When discussing human–machine interaction in the literature, this term refers to an interface between the machine (computer) and the user. Such an interface is intended to improve the operator's productivity, increase the reliability of task completion, etc. The quality of interfaces is assessed in terms of solving the above tasks for particular application fields. To improve interface quality, researchers and engineers typically rely on the psychological traits of human perception, motor characteristics, and other factors [2].

There are certain peculiarities in assessing the quality of human interaction with an LLM. First, it is a zero-interface system: a human communicates with the computer in natural language, e.g., by voice or via a standard ergonomic keyboard. Second, an LLM responds to a query in natural language (not as a pointer device or some other type of meter), and here, individual psychological traits must be considered. However, these are not the characteristics of perception or motor skills, but rather the mechanisms of thinking, understanding, reasoning, and so on.

Some publications were devoted to assessing the quality of explanations provided by AI systems, the so-called *eXplainable AI* (XAI) systems [3]. They explain the neural network's reasoning (inference) behind the response to a query. The explanation of inference is intended to increase confidence in an AI system and eliminate errors when making critical decisions. Such explanations include causal chains from a

query to the result in terms of concepts from the subject area where the neural network has been trained. An explanation contains no subject area semantics but only a logical inference.

Obtaining such an inference in an LLM is associated with certain challenges. An LLM is a deep learning neural network. As is well known, deep learning networks are data representation networks. This means that a trained neural network represents a data structure contained in a training dataset. LLMs are trained on a large text corpus, and built-in attention mechanisms highlight the syntactic structure of the text [4]. In other words, methods for extracting causal inference chains are pointless to apply to understand the correctness of the result. The inference will indicate the syntactic structure of the sentence, but not the specifics of the subject area.

The capabilities of an LLM in natural language processing, comprehension, and reasoning are assessed using special test programs (benchmarks). The same benchmarks can be applied to assess the human's capability of solving complex tasks presented in natural language and to compare the results of a human's solution with those of an LLM [5]. In this case, the capabilities of a human and an LLM are assessed separately using the same benchmarks, and the results are compared to judge the success of the model's developers and the extent to which their model approaches the human in terms of capabilities.

Benchmarks serve to test the individual capabilities of an LLM. As a rule, benchmarks have an unambiguous answer to a given test question. Therefore, a particular capability can be measured. When making decisions in complex social and economic situations, all these capabilities operate simultaneously in various combinations. Furthermore, in such complex situations, there exists a set of possible solutions, and it is difficult to choose the best one: all must be tested in practice or require additional assessment and justification. In such situations, conventional benchmarks do not work, and expert assessment is needed accordingly.

Its distinctive feature lies in the necessity to assess the interaction between a human and an LLM, which must be viewed as a complex cognitive system intended to generate high-quality solutions for managing complex situations. In this case, an LLM outputs a statistically plausible response (the most likely response included in the training sample) without understanding its meaning, while a human attempts to understand the response and integrate it into the individual knowledge system. The human's comprehension of the response provided by an LLM essentially depends on their level of knowledge; therefore, receiving the same response from an LLM, different people may assess it differently. In this case, the system for assessing the explanation must be personalized.

Thus, the problem is to develop an individual assessment method for the human's satisfaction with the explanations of the decision situation received from an LLM.

In part I of the study, the goals and tasks of explanation in decision support processes under uncertainty have been formulated [1]. Part II considers a method for assessing the satisfaction of the decision-maker (DM) with an explanation for solving investigation and practical (pragmatic) tasks. Based on the results of experiments with explanations from real LLMs, we divide these explanations into the classes of explanation models discussed in part I of the study [1].

1. QUALITY CRITERIA FOR EXPLANATIONS IN THE HUMAN–LLM SYSTEM

The systemic characteristics of an explanation system were presented in the paper [3].

First, an explanation system is a complex cognitive system comprising a human, their intellect, and a trained neural network. The human's mental space is engaged in the functioning of explanations. It is responsible for understanding the responses generated by an LLM. If the language model's explanation is understandable to a human, their mental space will expand.

Second, an explanation system is a dialog system. If an LLM answers a question, the user may have follow-up questions. In this case, a good approach for follow-up questions is Five Whys [6]. This method allows one to construct causal chains of a problem and identify its root cause.

Third, an explanation system is an investigation system. It should stimulate the human's curiosity, encouraging them to understand the problem and find the right solution.

In [7, 8], commonsense prompts for LLMs were investigated using benchmarks. In addition to improving the quality of the model's performance through such prompting, the human factor incorporated into a cognitive explanation system was studied therein. As noted, experts assess some correct explanations of an LLM (according to a benchmark) as incorrect, and, conversely, some incorrect ones as correct. Moreover, some correct answers are assessed as completely helpless. Hence, the quality of a cognitive DSS with a human and an LLM essentially depends on the former's level of knowledge and is determined by the human.

To assess the quality of such a cognitive system, the authors of [7, 8] conducted an expert survey based on the following criteria:

- Fluency: Is the explanation grammatically correct? (Scale values: grammatical; not entirely ungrammatical but understandable; completely not understandable).
- Relevance: Does the explanation correspond to the question's topic? (Scale values: relevant; irrelevant).
- Factual correctness: Is the explanation factually correct? (Scale values: factually correct; likely true).
- Helpfulness: Does the explanation help answer the question, directly or indirectly? (Scale values: helpful; neutral; harmful).

Obviously, these criteria scales are nominal. Working with such scales requires a large number of respondents and specific methods for processing survey results. Judging by the works [7, 8], the assessment was carried out for a small number of respondents, and an acceptable consistency in the answers was achieved.

For a prompt based on generating commonsense suggestions using an LLM, 82% of the explanations were grammatical and factually correct, while only 72% were helpful [7]. For a prompt based on an internal dialog with an LLM, only 40% of the factually correct answers were helpful [8].

According to the results of the respondent survey, it is important to assess the "human–LLM" cognitive system.

In [3], several criteria were proposed for assessing the quality of explanation systems:

– the helpfulness of an explanation,
– the understandability of an explanation,
– the implementability of an explanation,
– trust in an explanation,
– the ability of an explanation to stimulate curiosity.

These criteria can be measured. In [3], the following method was proposed to determine the criteria values in the complex cognitive system under consideration: a group of respondents tests the LLM's explanations for each criterion using the Likert ordinal scale [9].

The objective of the testing procedure is to measure a latent variable, i.e., the opinions of respondents regarding the quality of the LLM's explanation for the regularities of a subject area, and, if possible, to assess the expansion of the human's mental space owing to the explanations. In the sequel, we will utilize this system of criteria, referring to the authority of its developers as renowned experts in the field of human–computer interaction assessment.

2. MEASURING LATENT VARIABLES

In sociology and psychology, an expert survey is a conventional approach to measuring latent variables. A latent variable is understood as the attitude of a person or social group toward some directly unmeasurable phenomenon. A latent variable can be determined through certain presumably related variables measured on ordinal or interval scales. The value of a latent variable on such scales is determined using methods of mathematical statistics and measurement theory.

Numerous methods and techniques have been developed for measuring latent variables in sociology. We will not dwell on them here, merely explaining the core. To measure a latent variable through observable and measurable variables, the investigator develops special questionnaires for each measurable variable with possible answer options. The questionnaires are distributed to a large number of respondents, who answer the questions. Next, various methods of mathematical statistics are applied to process the survey results and assess the latent variable. It is desirable to assess a latent variable on an interval scale. In this case, it becomes possible to compare the values of latent variables across a set of measured objects and assign numerical assessments (scores) to such comparisons.

The challenge lies in obtaining an interval estimate for a latent variable. For this purpose, the group of respondents must be homogeneous. Special studies are conducted to form a group of respondents. Even if the group is selected, inhomogeneities may still arise in the survey, to be corrected by the investigator manually. This procedure is called sample cleaning; it often leads to the formation of a new group of respondents, changes to the questionnaire, or modifications to the answer options. The survey is repeated until acceptable results are obtained. Note that survey organization includes group formation, questionnaire development, processing of the results, and possibly sample cleaning or a repeated survey; all these steps require significant time and financial resources.

3. THE REASONABLE RESPONDENT MODEL AND GROUPS OF RESPONDENTS

The quality assessment of an LLM's explanation in decision support processes proposed in this paper has the following specifics: the human's understanding of the model's response much depends on their level of knowledge; therefore, different people may assess the same explanations from an LLM differently. In this case, the assessment model must be personalized.

Group assessments of the human's satisfaction with an LLM's explanation may be inaccurate.

Currently, the so-called model-based approach is popular in sociology [10]. With the development of AI technologies, a new scientific line—computational sociology—is emerging in the study of social systems. This is an interdisciplinary field that uses computational methods and big data analysis to study social phenomena, i.e., to develop and test theories about social systems, behavioral models, and interactions [11].

In this paper, we construct a reasonable respondent model, represent a set of reasonable respondents as a multi-agent system (MAS) simulating a homogeneous group of respondents, and subsequently measure a latent variable using conventional sociology methods.

The parameters of the reasonable respondent model depend on the set of measurable variables, the characteristics of the model's subject area, and the DM's preferences. Therefore, we will construct a model based on the previously defined criteria for assessing the DM's satisfaction with explanations. Recall that the criteria are the helpfulness of an explanation, the understandableness of an explanation, the implementability of an explanation, the ability of an explanation to stimulate curiosity, and trust in an explanation.

In the model-based approach to sociology, a model must reflect the regularities of respondent's behavior in the processes of assessing the measurable variables. These behavioral regularities are described using structural equations, i.e., the relationships between the values of measurable variables [12]. As a rule, these are linear equations of the general form $Y = X + \tilde{x}$, where Y is a latent variable; X is a measurable variable; and $\tilde{x}$ is a random measurement error. If the number of measurable variables is large, the relationships can be represented as a digraph for clarity. Figure 1 shows such a graph for the criteria under consideration.

In this case, the structural equations realizing the behavior of a reasonable respondent take the form of inequalities

$$X_4 \le X_1 + \tilde{x_1}, \quad (1)$$

$$X_4 \le X_2 + \tilde{x_2}, \quad (2)$$

$$X_4 \le X_3 + \tilde{x_3}, \quad (3)$$

$$X_5 \ge X_4 + \tilde{x_4}, \quad (4)$$

where X_1, X_2, X_3, and X_4 are the measurable variables; $\tilde{x_1}$, $\tilde{x_2}$, $\tilde{x_3}$, and $\tilde{x_4}$ are the random measurement errors of the variables.

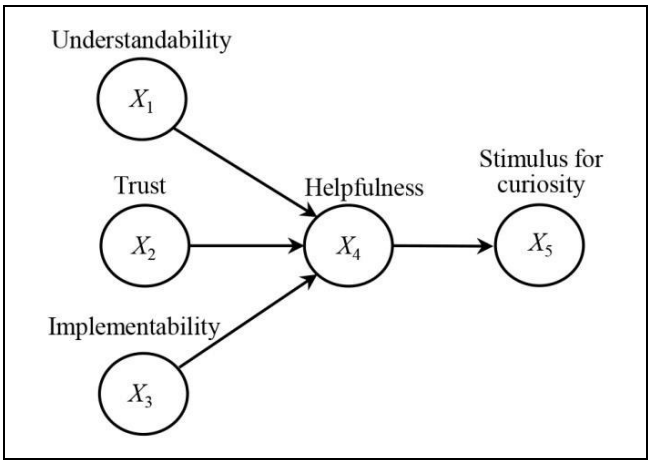

Fig. 1. The structural model of a reasonable respondent.

The values of these variables depend on the method of measurement. Let us define the Likert ordinal scale for measuring the values of the variables and, hence, for determining their possible values (Table 1).

Inequalities (1)–(4) embody the regularities of the respondent's reasonable behavior. Due to inequality (1), an explanation cannot be helpful without understandableness. By inequality (2), an explanation cannot be helpful without trust. Inequality (3) implies that an explanation cannot be helpful without implementability. According to inequality (4), if an explanation is helpful, the stimulus for curiosity decreases. This means that the explanation has been "integrated" into the human's knowledge system, expanding it, and further studies become unnecessary.

Note that the reasonable respondent model is constructed by the user of the LLM's explanation, i.e., it is subjective. The user selects the criteria—the measurable variables—and defines their relationships: the model described above can be restructured by the user to fit their beliefs. We also emphasize that the resulting subjective model may reflect the behavior of "like-minded" respondents with the same views as the DM (model designer). Here, the matter concerns a group of respondents sharing common opinions and beliefs and exhibiting similar behavior when assessing criteria in similar decision situations. If such a group of respondents is surveyed, the results of the survey will be homogeneous and can be used to assess satisfaction with LLM's explanations.

We tested the hypothesis regarding the homogeneity of the group of respondents acting in accordance with the reasonable behavior model during the assessment.

The Likert scale [9], a method popular in sociology, was used to assess the values of the measurable variables; see Table 1.

Table 1

The Likert scale

Helpfulness	Implementability	Understandability	Stimulus for curiosity	Trust	Score
Very helpful	Easy to implement	Absolutely understandable	Definitely arises	Completely trustworthy	7
Helpful	Implementable	Understandable	Arises	Trustworthy	6
Slightly helpful	Difficult to implement	Slightly understandable	Weakly arises	Weakly trustworthy	5
Unsure	Unsure	Unsure	Unsure	Unsure	4
Rather helpless	Rather non-implementable	Rather non-understandable	Rather does not arise	Rather untrustworthy	3
Helpless	Non-implementable	Non-understandable	Does not arise	Untrustworthy	2
Completely helpless	Impossible to implement	Completely non-understandable	Does not arise at all	Completely untrustworthy	1

The columns in Table 1 present the criteria and their possible linguistic values. The linguistic values correspond to numbers from 1 to 7 on an ordinal scale. When assessing an LLM's explanation, a respondent answers questions based on the five criteria, assigning the scores.

The respondent's scores are represented as a numerical vector $R = (r_1, r_2, r_3, r_4, r_5)$, where $r_i \in \{1, 2, 3, 4, 5, 6, 7\}$, and all scores must satisfy the reasonable respondent model.

An interesting hypothesis was put forward by L. Thurstone, a renowned psychophysicist and one of the founders of measurement theory in sociology [13]. If the same respondent is asked to assess a certain phenomenon or object at different times, the resulting temporally separated assessments will obey the Gaussian distribution with definite mean and variance; and if a group of respondents is asked to assess the same phenomenon or object simultaneously, their assessments will also be distributed according to the Gaussian law with definite mean and variance. Under certain conditions, the Gaussian distributions of a single respondent and a group may coincide. (For example, if the group of respondents is homogeneous and all group members share the views of the respondent with temporally separate assessments.) Thurstone's rigorous mathematical proof of the measurability of a latent variable is based on this hypothesis and several other assumptions [13, 10].

Suppose that a set G of respondents has be surveyed to obtain a set $R_j = (r_{1j}, r_{2j}, r_{3j}, r_{4j}, r_{5j})$ of scores for an observed phenomenon or object; here, $r_{1j},\ldots, r_{5j}$ are the scores of the jth respondent for all five criteria of the Likert scale $r_{ij} \in \{1, 2, 3, 4, 5, 6, 7\}$, $i \in \{1, 2, 3, 4, 5\}$ denotes the criterion number, and $R_j \in G$.

For the group of respondents G, we define the mean and standard deviation for each criterion as the vector of pairs $(M(r_1), \sigma_1; M(r_2), \sigma_2; M(r_3), \sigma_3; M(r_4), \sigma_4; M(r_5), \sigma_5)$, where $M(r_i)$ and σ_{ij} are the mean and standard deviation, respectively, of the group's assessment according to the ith criterion.

The mean vector $(M(r_1), M(r_2), M(r_3), M(r_4), M(r_5))$ is supposed to characterize the latent variable (satisfaction with an LLM's explanation). However, the scores of the group of respondents must be consistent. Consistency is checked using Cronbach's α [14]. A value of this coefficient below 0.5 indicates that the scores within the group are inconsistent and the mean vector does not characterize the latent variable. It is necessary to discard contradictory answers and recheck the consistency of the scores, or assemble a new group of respondents, or modify the questions, conduct a new survey, and re-determine the consistency of the scores. This procedure is repeated until an acceptable value of the consistency coefficient is obtained.

Let us construct a model for a group of like-minded respondents. Consider a respondent R who has provided an individual score $R = (r_1, r_2, r_3, r_4, r_5)$, where $r_i \in \{1, 2, 3, 4, 5, 6, 7\}$ and $i \in \{1, 2, 3, 4, 5\}$. This respondent (and their score R) will be called basic.

Using the basic respondent's scores, we form the set of possible scores according to the following rule: $R_j = (r_1 + \tilde{x_{1j}}, r_2 + \tilde{x_{2j}}, r_3 + \tilde{x_{3j}}, r_4 + \tilde{x_{4j}}, r_5 + \tilde{x_{5j}})$, where $r_1,\ldots, r_5$ are the basic respondent's scores for the criteria; $\tilde{x_{ij}}$ is the random error in the jth respondent's score (the estimation error) for the ith criterion. Assume that the respondent's random error is distributed according to the Gaussian law: $\tilde{x_{ij}} \in [-\sigma, +\sigma]$, where σ is the standard deviation of the estimation error.

For the experiment, we applied Excel's random number generator, i.e., the function *NORM.INV*(*RAND*(); r_i; σ), where r_i is the basic respondent's score for the *i*th criterion; σ is the standard deviation. As a result, 1000 hypothetical surveys were formed, including 600 hypothetical respondents with the standard deviations of the estimation error σ = 2 and σ = 3. In each case under consideration (1000 surveys of 600 respondents for σ = 2, and 1000 surveys of 600 respondents for σ = 3), Cronbach's α (5) was calculated [14]; in all the cases, it indicated unsatisfactory survey consistency with values below 0.5 and even negative values.

Next, from the hypothetical surveys, we selected only those questionnaires that satisfy the respondent's reasonableness inequalities (1)–(4). Under different random generations of the scores, between 40 and 70 (6–12%) questionnaires out of the 600 random ones were reasonable (see the above rule). These questionnaires form a homogeneous group of reasonable respondents.

To verify the consistency of this group's scores, we used Cronbach's α [14]

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_x^2}\right), \qquad (5)$$

where k is the number of questions asked to respondents; σ_i^2 is the standard deviation of the respondent's scores for each question; and σ_x^2 is the standard deviation of the sum of the respondents' scores.

Among the 1000 surveys, from which the questionnaires of the basic respondent's like-minded ones were selected, the Cronbach's α exceeded 0.7 for 97% of the groups, indicating good consistency.

Cronbach's α was tested for the questionnaires generated for two values of the standard deviation of respondents' scores from the mean: σ = 2 and σ = 3. The results of the experiment are presented in Fig. 2.

Obviously, for different values of the standard deviation, the number of consistent surveys changes insignificantly; approximately 97% of the surveys are consistent. Therefore, the respondent groups based on the reasonable respondent model are homogeneous.

The criteria can be used to measure the value of a latent variable if they are correlated with one another.

Let us define Spearman's rank correlation coefficient for the criteria by the formula

$$\rho = 1 - 6\frac{\sum_{i=1}^{n} d_i^2}{n^3 - n}, \qquad (6)$$

where n is the number of respondents; d_i^2 is the squared difference between the ranks of respondents' scores for each question and the ranks of the sum of these scores.

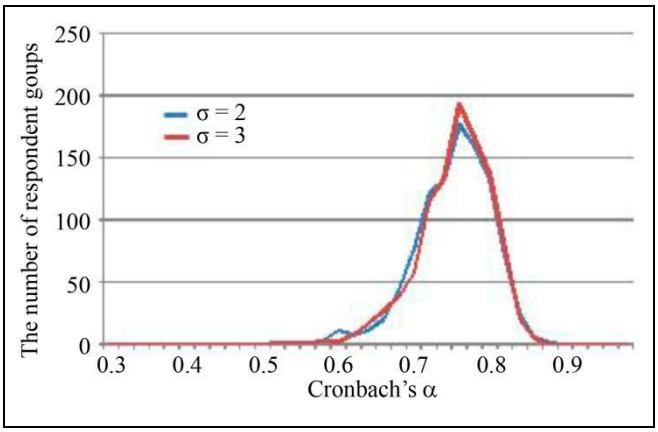

Fig. 2. Cronbach's α for σ = 2 and σ = 3.

Figure 3 shows the coefficient (6) for the 1000 surveys of the groups of like-minded respondents.

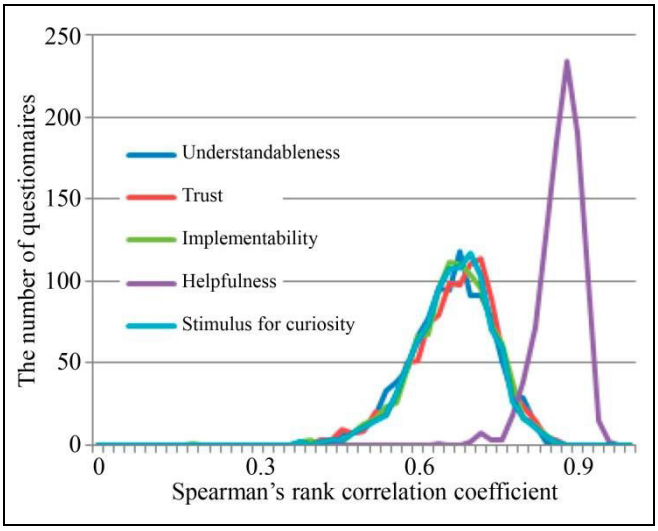

Fig. 3. Spearman's rank correlation coefficient for all criteria.

According to the graph, 97% of the questionnaires have a rank correlation above 0.5.

For each correlation coefficient, we determine its statistical significance. The observed value of the *t*-test is given by

$$t_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad (7)$$

where n is the number of respondents; r is the correlation coefficient, $r \in R$.

Using the table of critical values of Student's *t*-test, we determine the critical value t_{cr} for n. If the ob-

served value (7) exceeds the critical one ($t_{obs} > t_{cr}$), then the correlation is significant; in other words, the correlation hypothesis can be accepted.

The critical values of the t-test for $n = 40$ and significance levels of 0.05 and 0.01 are $t_{cr(0.05)} = 2.02$ and $t_{cr(0.01)} = 3.55$, respectively; for $n = 70$, they are $t_{cr(0.05)} = 2.0$ and $t_{cr(0.01)} = 3.46$. Judging by Fig. 4, for 99% of the questionnaires, the observed value of the t-test is higher than the critical ones. Consequently, the correlation coefficients for all assessed criteria are statistically significant.

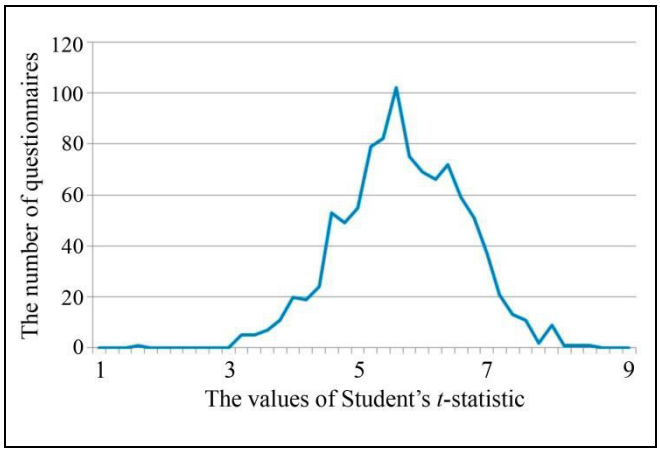

Fig. 4. The observed values of Student's t-statistic.

Note that the results of real field studies with the selection of a group of respondents and their survey undergo all of the statistical tests above. In this case, the values of the consistency, rank correlation, and statistical significance indices obtained for the hypothetical group are assessed as high-quality. Moreover, depending on the survey's objective and subject area, the reasonable respondent model may differ. However, if an investigator has constructed and justified such a model, it can be useful when processing the results of sociological or psychological field studies. Such a model can serve as a filter during the sample cleaning stage (exclusion of low-quality questionnaires). Typically, when cleaning a sample, an expert excludes invalid questionnaires based on their own subjective model of respondent's reasonableness. In this case, the cleaned sample will reflect the expert's subjective model of reasonableness. If this subjective model is formalized, the sample cleaning process can be automated. As shown by the above experiments, consistent questionnaires with estimates of the latent variable can be obtained from randomly generated survey scores using the reasonable respondent model.

Thus, a single respondent's assessment of an LLM's explanation, based on the model of their reasonable behavior, reflects the behavior of a homogeneous group of like-minded respondents and can be used to select the best LLM as a DM's assistant.

4. ASSESSMENT OF THE DM'S SATISFACTION WITH AN LLM'S EXPLANATION

Satisfaction with an explanation was assessed on the Likert scale across five parameters (the score vector). A multicriteria assessment can be reduced to a single numerical score. Multicriteria assessment methods are widely covered in the literature. For example, the simplest method is the linear convolution: for each criterion, significance coefficients are determined, and the products of the criterion scores and the corresponding significance coefficients (weights) are summed up. However, this method suffers from some drawbacks. In particular, the summation of weighted scores leads to a compensation effect: a low score on an important criterion can be offset by the sum of high scores on unimportant criteria.

Therefore, we apply an expertise method to determine two scores characterizing ideal explanations for investigation and practical (pragmatic) tasks, respectively [1]. For the expert assessment of ideal explanations, we take a Likert scale; see the scores in Table 2.

Table 2

The scores of ideal explanations for investigation and pragmatic tasks

Task type	Understandableness	Trust	Implementability	Helpfulness	Stimulus for curiosity
Investigation task	Completely understandable (7)	Rather untrustworthy (3)	Rather non-implementable (3)	Very helpful (7)	Definitely arises (7)
Pragmatic task	Completely understandable (7)	Completely trustworthy (7)	Easy to implement (7)	Very helpful (7)	Rather does not arise (3)

Let us explain these scores of ideal explanations.

To solve an investigation task, an explanation must be completely understandable, very helpful, and must provide a stimulus for further investigation, while factual correctness and implementability are not as important. Note that in heuristic solution generation methods, counterfactual and non-implementable alternatives are considered rather than discarded. They are supposed to stimulate the development of creative solutions and further analysis. Therefore, factual correctness and implementability have low priority for an investigation task. Here, the search is for appropriate explanations that should ultimately justify and implement the solution itself.

To solve a pragmatic task, an explanation must be completely understandable, very helpful, factually correct, and implementable. At the same time, the stimulus for further investigation should diminish: the search is complete, and one proceeds to implementation.

An investigation task is addressed until easily implementable and factually correct solutions are found. In this case, one approaches the ideal point for a pragmatic task.

An ideal explanation for solving an investigation task is the score vector $R_{inv} = (r_{inv1}, r_{inv2}, r_{inv3}, r_{inv4}, r_{inv5}) = (7, 3, 3, 7, 7)$. An ideal explanation for solving a pragmatic task is the score vector $R_{prg} = (r_{prg1}, r_{prg2}, r_{prg3}, r_{prg4}, r_{prg5}) = (7, 7, 7, 7, 3)$.

All possible scores across all criteria are the set of all possible score vectors from the direct product of the Likert scale, $SR = \underset{i}{\times} E_i$, where $E_i = \{1, 2, 3, 4, 5\}$; i denotes the criterion number, $i \in \{1, 2, 3, 4, 5\}$. In other words, $R_{prg}, R_{inv} \in SR$. Similarly, all scores assigned to explanations by the jth respondent, $R_j = (r_{1j}, r_{2j}, r_{3j}, r_{4j}, r_{5j})$, $r_{ij} \in \{1, 2, 3, 4, 5, 6, 7\}$, belong to the direct product of Likert scales: $R_j \in SR \ \forall j$.

We define the distance between the score vector R_j of an LLM's explanation for an investigation task and the score vector R_{invj} of the ideal explanation for the jth question as $St_{invj} = \rho(R_j, R_{inv})$, where ρ is a distance metric, and $St_{invj} \in [0, \rho(R_0, R_{inv})]$.

By analogy, the distance between the score vector R_j of an LLM's explanation for a pragmatic task and the score vector R_{prgj} of the ideal explanation for the jth question is defined as $St_{prgj} = \rho(R_j, R_{prg})$, where ρ is a distance metric, and $St_{prgj} \in [0, \rho(R_0, R_{prg})]$.

Here, $R_0 \in SR$ is the vector of the minimum elements of the set SR, i.e., $R_0 = (1, 1, 1, 1, 1)$. In the case under consideration, $\rho(R_0, R_{inv}) = 22$ for an investigation task, and $\rho(R_0, R_{prg}) = 26$ for a pragmatic task.

Let the distance metric be Minkowski's "urban neighborhood" metric. Then, the distance St_{invj} for an investigation task is given by

$$St_{invj} = \sum_{i=1}^{5} |r_{ij} - r_{invj}|, \tag{8}$$

where j and i denote the question number and the criterion number, respectively.

The distance St_{prgj} for a pragmatic task is given by

$$St_{prgj} = \sum_{i=1}^{5} |r_{ij} - r_{prgj}|. \tag{9}$$

It is convenient to assess the LLM's achievement of an ideal explanation in percentage terms.

Definition 1. The degree $St^{\%}_{invj}$ of the DM's satisfaction with an LLM's explanation for an investigation task is the degree to which the user's score vector for the jth explanation achieves an ideal explanation of this task, expressed as a percentage:

$$St^{\%}_{invj} = \left(1 - \frac{St_{invj}}{\rho(R_0, R_{inv})}\right) \cdot 100\%. \tag{10}$$

Definition 2. The degree $St^{\%}_{prgj}$ of the DM's satisfaction with an LLM's explanation for a pragmatic task is the degree to which the user's score vector for the jth explanation achieves an ideal explanation of this task, expressed as a percentage:

$$St^{\%}_{prgj} = \left(1 - \frac{St_{prgj}}{\rho(R_0, R_{prg})}\right) \cdot 100\%. \tag{11}$$

Formulas (10) and (11) can be used to calculate and compare the DM's satisfaction degrees on various questions and for different LLMs. Hence, it is possible to select the best LLM that meets the DM's needs in decision support processes.

5. EXAMPLE

To illustrate the use of an LLM as a DM's assistant, two Russian LLMs were considered: GigaChat 2.0 from Sberbank and YandexGPT 5 Pro from Yandex. Both suppliers claim remarkable capabilities of their LLMs, offering a free version and the option to purchase a license for additional features. We applied the free versions of the LLMs, available on the official websites [15, 16]. These models include the developers' default prompts. User-defined prompts were not used. This example is intended to demonstrate the novel method for assessing user's satisfaction with explanations provided by the two LLMs.

Both models were asked the same "why?" question. The response from GigaChat 2.0 was used to generate the next "why?" question until the problem was understood. Thus,

Five Whys for identifying causes was implemented. This is the investigation-oriented phase of the decision process, which creates an information environment to support decision-making through LLM's explanations. It was followed by the practical phase, in which the DM attempts to obtain additional explanations from an LLM regarding the implementability of an alternative under consideration. In this case, an LLM is asked questions such as "how?" or "what kind of?"

The LLMs responded with detailed answers to the questions; see the Appendix below. (Note that the questions and the models' responses were formulated in the Russian language; see the original Russian text in *Problemy Upravleniya*, 2026, no. 2, pp. 75–77. The questions and responses provided in the Appendix of this paper are a translation of the original ones into English, for the reader's convenience.) The comments on the LLM's answers justify the expert scores based on five criteria.

Question 1: "Why do businesspeople encounter financial problems?" The detailed answers from the two LLMs can be found in Table A1 of the Appendix.

Comment. The response from GigaChat 2.0 includes a summary of the answer and the following seven causes of the problem: insufficient cost control; lack of capital; noncompliance with tax laws; poor risk management; low profitability of products or services; flawed marketing strategy; and personnel and work organization. Each cause is described using a template: a brief characterization of the cause, possible consequences, and possible actions to mitigate the negative consequences.

The scores of GigaChat 2.0's explanation for Question 1 are presented in Table 3. The developers of this model do not reveal their professional secrets; as it seems, few-shot prompting is embedded into the model.

The explanation provided by YandexGPT 5 Pro lists two main causes of financial problems: insufficient planning and market analysis; and ineffective financial management. Each cause is explained; typically, the possible consequences and difficulties due to these causes are listed. This explanation is not completely understandable to an expert. The expert score for this model is lower than for GigaChat 2.0 (Table 3).

GigaChat 2.0's scores for Question 1 are $R_1^{GigaChat} = (7, 6, 5, 5, 6)$.

YandexGPT 5 Pro's scores for Question 1 are $R_1^{GPT\ 5\ Pro} = (6, 5, 5, 5, 7)$.

Question 2 (a pragmatic "how?") was formulated based on the causes of financial problems described by GigaChat 2.0. The question is: "How can the effectiveness of expense control be improved?"

The detailed explanations of the two LLMs are provided in Table A2 of the Appendix.

Comment. GigaChat 2.0's explanation contains two methods for improving expense control: budgeting and automation of expense tracking and analysis. For each method, the explanation is presented in the form of a template (a description of the method and its advantages). This facilitates understanding of the explanation. Such an explanation appears correct but somewhat abstract, which is more suitable for elaborating a business development strategy. Therefore, the score for the implementability criterion was slightly reduced. The scores can be found in Table 4.

In its response, YandexGPT 5 Pro also describes two methods for solving financial problems: the implementation of a budgeting and financial planning system; and the analysis and optimization of the expense structure. The actions for implementing these methods are characterized as well. Therefore, the explanation is more suitable for making decisions on operational business management. Here, the implementability scores for the explanation are high (Table 4).

GigaChat 2.0's scores for Question 2 are $R_2^{GigaChat} = (7, 6, 6, 6, 6)$.

YandexGPT 5 Pro's scores for Question 2 are $R_2^{GPT\ 5\ Pro} = (6, 5, 7, 6, 6)$.

Table 3

The scores of LLM's explanations for Question 1

Model	Understandableness	Trust	Implementability	Helpfulness	Stimulus for curiosity
GigaChat 2.0	Completely understandable (7)	Trustworthy (6)	Difficult to implement (5)	Slightly helpful (5)	Arises (6)
YandexGPT 5 Pro	Understandable (6)	Weakly trustworthy (5)	Difficult to implement (5)	Slightly helpful (5)	Necessarily arises (7)

Table 4

The scores of LLM's explanations for Question 2

Model	Understandableness	Trust	Implementability	Helpfulness	Stimulus for curiosity
GigaChat 2.0	Completely understandable (7)	Trustworthy (6)	Implementable (6)	Helpful (6)	Arises (6)
YandexGPT 5 Pro	Understandable (6)	Weakly trustworthy (5)	Easy to implement (7)	Helpful (6)	Arises (6)

Question 3 (pragmatic) aims to justify the use of "budgeting" when solving financial problems in business. The question is as follows: "How can budgeting be implemented in a small-sized business? List two methods." The detailed responses of the models are presented in Table A3 of the Appendix.

Comment. GigaChat 2.0's explanation provides two implementation methods: top-down and bottom-up. As before, the explanation of this LLM is in a template that includes a description of the method and its advantages and drawbacks. The explanation is understandable and helpful; the score for the implementability criterion is somewhat low. The scores are combined in Table 5.

The explanation provided by YandexGPT 5 Pro also contains a description of two methods: creating a simple and transparent budget with main revenue and expense items, and appointing a person responsible for budgeting and budgetary supervision. Each method is explained. The practical focus of these methods is evident from their names in the LLM. In other words, the explanations of these methods have practical tips for implementing budgeting in business. The scores for this model's explanation can be found in Table 5.

GigaChat 2.0's scores for Question 3 are $R_3^{GigaChat} = (7, 7, 6, 6, 5)$.

YandexGPT 5 Pro scores for Question 3 are $R_3^{GPT\ 5\ Pro} = (6, 6, 7, 6, 5)$.

Question 4 (pragmatic) aims to explain the cost of implementing budgeting. The question is as follows: "What is the cost of implementing budgeting in a small-sized business?" The detailed responses of the models are provided in Table A4 in the Appendix.

Comment. GigaChat 2.0's explanation outlines the cost structure of implementing budgeting in a business. These are the main costs for software, consulting services, staff training, preparation of regulations and internal documents, and a general cost estimate. For each cost item, explanations and recommendations are provided regarding the selection of software products and their cost, the cost of consulting and staff training, etc. The price range for implementing budgeting is given. Of course, all recommendations and prices require verification, but overall, the scores for the explanation based on this criterion are high (see Table 6).

The response from YandexGPT 5 Pro contains tips and recommendations for determining the cost of implementing budgeting in a business, depending on the size of the business and the skill level of the staff. The explanation includes possible cost items without specifying their cost. The scores for this model's explanation are also presented in Table 6.

GigaChat 2.0's scores for Question 4 are $R_4^{GigaChat} = (7, 5, 6, 6, 6)$.

YandexGPT 5 Pro's scores for Question 4 are $R_4^{GPT\ 5\ Pro} = (6, 5, 7, 5, 6)$. ♦

6. EXPLANATIONS PROVIDED BY GIGACHAT 2.0 AND YANDEXGPT 5 PRO: ANALYSIS OF SATISFACTION DEGREES

Table 7 presents the expert assessments for the quality of explanations provided by GigaChat 2.0 for the four questions. For each question, the rows in the table show the expert assessments based on the criteria selected in Section 1. The last three columns contain the following data: the values of the distance St_{invj} and, separated by a slash, the expert's degree of satisfaction $St^{\%}_{invj}$ with the LLM's explanation when solving the investigation task; the values of the distance St_{prgj} and, separated by a slash, the expert's degree of satisfaction $St^{\%}_{prgj}$ with the LLM's explanation when solving the pragmatic task; finally, the difference $St^{\%}_{prgj} - St^{\%}_{invj}$, intended to assess the satisfaction with the explanation when solving the investigation and pragmatic tasks.

Table 5

The scores of LLM's explanations for Question 3

Model	Understandableness	Trust	Implementability	Helpfulness	Stimulus for curiosity
GigaChat 2.0	Completely understandable (7)	Completely trustworthy (7)	Implementable (6)	Helpful (6)	Weakly arises (5)
YandexGPT 5 Pro	Understandable (6)	Trustworthy (6)	Easy to implement (7)	Helpful (6)	Weakly arises (5)

Table 6

The scores of LLM's explanations for Question 4

Model	Understandableness	Trust	Implementability	Helpfulness	Stimulus for curiosity
GigaChat 2.0	Completely understandable (7)	Weakly trustworthy (5)	Implementable (6)	Helpful (6)	Arises (6)
YandexGPT 5 Pro	Understandable (6)	Weakly trustworthy (5)	Easy to implement (7)	Slightly helpful (5)	Arises (8)

The scores of one respondent for GigaChat 2.0's explanations match the reasonable behavior model. Table 7 presents the values of Cronbach's α for the responses to each question for a hypothetical group of like-minded individuals and the rank correlation coefficients for each criterion.

According to Table 7, GigaChat 2.0's explanations are closer to the ideal explanation of the pragmatic task than to that of the investigation task. To be fair, only Question 1 ("why?") relates to solving an investigation task, while the other questions ("how?" and "which?") relate to solving a pragmatic task. Based on Table 7, satisfaction with the explanations for Questions 2 and 3 when solving the pragmatic task increased from 76.9% to 84.6%. This may indicate an expansion of the expert's information environment for decision-making. However, the explanation in response to Question 4 reduced satisfaction to 71.1%. This is a normal phenomenon in the investigation process of finding solutions.

The scores of one respondent for YandexGPT 5 Pro's explanations match the reasonable behavior model. Table 8 presents the values of Cronbach's α for the responses to each question for a hypothetical group of like-minded individuals and the rank correlation coefficients for each criterion.

Similar to Table 7, the rows in Table 8 show the expert assessments based on the criteria selected in Section 1. By analogy, the last three columns contain the following data: the values of the distance St_{invj} and the expert's degree of satisfaction $St^{\%}_{invj}$ with the LLM's explanation when solving the investigation task; the values of the distance St_{prgj} and, separated by a slash, the expert's degree of satisfaction $St^{\%}_{prgj}$ with the LLM's explanation when solving the pragmatic task; finally, the difference $St^{\%}_{prgj} - St^{\%}_{invj}$, intended to assess the satisfaction with the explanation when solving the investigation and pragmatic tasks.

Table 7

The scores on different criteria and the degree of satisfaction with GigaChat 2.0's explanations

Question / Cronbach's α	Understandableness / Correlation coefficient	Trust / Correlation coefficient	Implementability / Correlation coefficient	Helpfulness / Correlation coefficient	Stimulus for curiosity / Correlation coefficient	Investigation task. Distance/ The degree of satisfaction $St^{\%}_{invj}$	Investigation task. Distance/ The degree of satisfaction $St^{\%}_{prgj}$	$St^{\%}_{prgj} - St^{\%}_{invj}$
Question 1 α = 0.80	7 ρ = 0.62	6 ρ = 0.80	5 ρ = 0.77	5 ρ = 0.86	6 ρ = 0.67	8 63.6%	8 69.2%	5.6%
Question 2 α = 0.82	7 ρ = 0.76	6 ρ = 0.68	6 ρ = 0.67	6 ρ = 0.90	6 ρ = 0.67	8 63.6%	6 76.9%	13.3%
Question 3 α = 0.81	7 ρ = 0.75	7 ρ = 0.77	6 ρ = 0.65	6 ρ = 0.90	5 ρ = 0.76	10 54.5%	4 84.6%	30.1%
Question 4 α = 0.79	7 ρ = 0.73	5 ρ = 0.70	6 ρ = 0.83	6 ρ = 0.84	6 ρ = 0.70	7 62.2%	7 71.1%	4.9%

Table 8

The scores on different criteria and the degree of satisfaction with YandexGPT 5 Pro's explanations

Question / Cronbach's α	Understandableness / Correlation coefficient	Trust / Correlation coefficient	Implementability / Correlation coefficient	Helpfulness / Correlation coefficient	Stimulus for curiosity / Correlation coefficient	Investigation task. Distance/ The degree of satisfaction $St^{\%}_{invj}$	Investigation task. Distance/ The degree of satisfaction $St^{\%}_{prgj}$	$St^{\%}_{prgj} - St^{\%}_{invj}$
Question 1 α = 0.71	6 ρ = 0.56	5 ρ = 0.73	5 ρ = 0.57	5 ρ = 0.87	7 ρ = 0.60	7 68.2%	11 57.7%	– 10.5%
Question 2 α = 0.80	6 ρ = 0.65	5 ρ = 0.73	7 ρ = 0.66	6 ρ = 0.87	6 ρ = 0.73	9 59.1%	7 73.1%	14.0%
Question 3 α = 0.82	6 ρ = 0.67	6 ρ = 0.70	7 ρ = 0.79	6 ρ = 0.80	5 ρ = 0.71	11 50.0%	5 80.8%	30.8%
Question 4 α = 0.71	6 ρ = 0.58	5 ρ = 0.74	7 ρ = 0.59	5 ρ = 0.85	6 ρ = 0.61	10 54.5%	8 69.2%	14.7%

According to Table 8, YandexGPT 5 Pro's explanations are closer to the ideal explanation for the pragmatic task than to that of the investigation task, except Question 1. Satisfaction with the explanations for Questions 2 and 3 when solving the pragmatic task increased from 73.1% to 80.8%. This indicates an expansion of the expert's information environment for decision-making; but the explanation for Question 4 reduced satisfaction to 69.2%.

Now we compare the expert's degrees of satisfaction with the explanations provided by GigaChat 2.0 and YandexGPT 5 Pro separately, based on satisfaction with solving investigation and pragmatic tasks. The comparison results are presented in Table 9.

Here, the rows contain the question numbers and the expert's degrees of satisfaction with the explanations provided by GigaChat 2.0 and YandexGPT 5 Pro when solving the investigation or pragmatic task. The difference $St^{\%}_{invj\_GigaChat} - St^{\%}_{invj\_GPT\ 5\ Pro}$ in the third column shows the distinction in the expert's satisfaction with the explanations when solving the investigation task. For Questions 2–4, GigaChat 2.0 has an obvious advantage. The difference $St^{\%}_{invj\_GigaChat} - St^{\%}_{invj\_GPT\ 5\ Pro}$ in the sixth column shows the distinction in the expert's satisfaction with the explanations when solving the pragmatic task. For Questions 1–4, GigaChat 2.0 also has a slight advantage.

According to Table 9, for Question 1 ("Why do businesspeople encounter financial problems?"), YandexGPT 5 Pro's explanation has a slight advantage (68.2%) over the counterpart provided by GigaChat 2.0 (63.6%) from the perspective of achieving the investigation purpose. At the same time, for achieving the pragmatic purpose, the explanation for this question from GigaChat 2.0 proved to be better (69.2%) than that of YandexGPT 5 Pro (57.7%). For Question 4 ("What is the cost of implementing budgeting in a small business?"), the explanation provided by GigaChat 2.0 for the investigation purpose of decision support proved to be better (68.2%) than that of YandexGPT 5 Pro (54.5%). The scores of the explanations for Questions 2 and 3 for investigation and pragmatic tasks slightly differ in favor of GigaChat 2.0.

Note that the analysis and assessment results presented here reflect the quality of a complex cognitive system comprising the DM and the LLM. These assessments depend on the LLM's capabilities and, moreover, on the expert's level of knowledge. Another expert or group of experts may arrive at different assessments and have different preferences when selecting an appropriate LLM as a decision-making assistant.

7. DISCUSSION

In part I of the study [1], the following classes of explanation models have been identified: the deductive-nomological model, unificationist models, and pragmatic models. They have been studied by philosophers; of course, there exists no generally accepted model of scientific explanation, and the debates continue. However, the main properties of these classes and their helpfulness for decision support are clear; for details, see part I of the study. An LLM generates explanatory text. May this text be considered an explanation from the standpoint of research methodology in the context of explaining phenomena, or may it not?

Let us analyze the relationship between the explanations generated by LLMs and the explanation models proposed within research methodology.

The explanations provided by GigaChat 2.0 are well-structured and follow a certain template. They are understandable and complete, but abstract. In research methodology, there are unificationist models [1, 17] that construct explanations according to a template. These models involve no logical inference, and therefore, it is difficult to verify the factual correctness of their explanations. However, an explanation in a unificationist model of explanation is derived based on an explanatory resource that contains verified and factually correct information. If GigaChat 2.0 uses a neural network trained on reference books and encyclopedias as its explanatory resource, trust in the explanations of this model will be high and useful for expanding the DM's information environment.

Table 9

Comparison of the expert's satisfaction with LLM's explanations when solving investigation and pragmatic tasks

Question	Investigation task			Pragmatic task		
	$St^{\%}_{invj\_GigaChat}$, %	$St^{\%}_{invj\_GPT\ 5\ Pro}$, %	$St^{\%}_{invj\_GigaChat} - St^{\%}_{invj\_GPT\ 5\ Pro}$, %	$St^{\%}_{prgj\_GigaChat}$, %	$St^{\%}_{prgj\_GPT5\ Pro}$, %	$St^{\%}_{prgj\_GigaChat} - St^{\%}_{prgj\_GPT\ 5\ Pro}$, %
1	63.6	68.2	–4.5	69.2	57.7	11.5
2	63.6	59.1	4.5	76.9	73.1	3.8
3	54.5	50.0	4.5	84.6	80.8	3.8
4	68.2	54.5	13.6	73.1	69.2	3.8

The explanations from YandexGPT 5 Pro are structured and understandable. Here, the explanations have a practical focus, as if they were trained on the experience of good and successful managers or consultants. In research methodology, there are pragmatic models of explanation [18]; see the description in part I of the study [1]. They are based on human experience and empirical observable data. The explanations of such models are characterized by empirical adequacy [18]. The acceptance of such explanations is based on trust, confidence that they are correct. YandexGPT 5 Pro can be associated with a model that generates pragmatic explanations. Trust in such explanations will be lower. It is necessary to consider the possible biases of the experts whose knowledge was used to train or fine-tune the model.

Thus, the two LLMs can be categorized, based on expert assessment, into different classes of explanatory models within research methodology. The classification of LLM's explanations allows one to assess, in the most general terms, the validity of these explanations and the challenges and prospects for their development.

The deductive-nomological model of explanation [1, 19] has been considered in part I of the study. This classical model of scientific explanation is based on deductive reasoning with known laws. No strict logical inference based on laws was found in the above explanations. As an example, the LLMs under consideration were asked the following question on economic laws: "Why do businesspeople face financial problems? Cite an economic law." The models' explanations are presented in Table A5 of the Appendix.

Obviously, both models refer to the same economic law—the law of matching revenues to expenses—and give the main causes for its violation. However, there is no strict logical inference here. Such an inference can be provided by reasoning LLMs, which are not considered here.

CONCLUSIONS

This paper has considered the issues of measuring the quality of a hybrid DSS composed of a DM and an LLM as an assistant. We have proposed a set of criteria for assessing the quality of the hybrid DSS and an expert assessment procedure. A distinctive feature of this procedure is the structural model of a reasonable expert. As shown, the reasonable respondent model allows creating a virtual group of respondents possessing the properties of consistency and homogeneity. The scores based on the reasonable respondent model can be used as quality assessments for an LLM's explanation. The degree of expert's satisfaction with an LLM's explanation when solving investigation and pragmatic tasks has been formally defined.

The operation and assessment of a DSS that includes an expert and GigaChat 2.0 and YandexGPT 5 Pro LLMs have been considered as an example.

According to the analysis of the response provided by the above models to expert's questions, their explanations can be classified as unificationist (GigaChat 2.0) and pragmatic (YandexGPT 5 Pro). The explanations of a unificationist model are based on the so-called explanatory reserve (e.g., verified information from reference books and encyclopedias). Such explanations are trustworthy, but they are strategic in nature and difficult to adapt for solving operational tasks. The explanations of a pragmatic model involve the experience of managers and experts, as well as empirical data. They are useful for solving operational tasks, but may include expert biases and lead to errors. The explanations of these two models are focused on expanding the informational environment for decision-making and generating alternatives.

A deductive-nomological model of explanations, based on logical inference and known laws, is suitable for justifying alternative solutions. The LLMs considered above provided no logical justification for alternative solutions. The class of reasoning LLMs and special prompting techniques are required to obtain a deductive-nomological model of explanation.

Further research into DSSs that include a DM and an LLM is envisioned as follows:

- development of prompts to support explanations for solving an investigation task using heuristic methods for structuring information, strategic planning, and fostering critical thinking and creativity (SWOT analysis, Ishikawa's diagrams, and others);
- development of prompts for implementing a deductive-nomological model of explanation.

The proposed method for measuring quality and assessing the DM's satisfaction, based on the structural model of a reasonable expert and quasi-experimental assessment, can be useful during the development of a DSS that includes an LLM. This method yields qualitative assessments for hybrid DSSs and reduces the time and cost of assessment without involving expert groups.

APPENDIX

Table A1

The explanations in response to the question

"Why do businesspeople encounter financial problems?"

GigaChat 2.0	YandexGPT 5 Pro
Internal factors 1. Insufficient expense control Many entrepreneurs face problems due to the absence of effective expense control. This includes excessive spending on unnecessary assets, inefficient operations, and a lack of financial statement analysis. It is important to regularly analyze expenses and look for opportunities to optimize costs. 2. Lack of capital Insufficient business financing is also a common cause of financial problems. A lack of sufficient startup capital or the inability to attract additional investment can lead to a shortage of funds to cover current obligations and expand the business. 3. Non-compliance with tax laws Tax compliance is crucial for any entrepreneur. Violating tax regulations or improper accounting can result in fines and additional financial obligations, which exacerbate the financial situation. 4. Poor risk management The absence of a risk management strategy can lead to significant losses. For example, unforeseen market changes, currency fluctuations, or the loss of key customers can negatively impact the company's financial position. 5. Low profit margins on products or services If a business offers products or services with low profit margins, this can become a source of ongoing financial problems. Entrepreneurs must thoroughly monitor their product line and respond promptly to changes in demand and competition. 6. Flawed marketing strategy Incorrect assessment of the market and consumer preferences can lead to losses. An effective marketing strategy shall consider the needs of the target audience and adapt to changes in the market environment. 7. Staff and work organization Personnel issues, such as low employee productivity, high turnover, or insufficient employee qualifications, can significantly impact business profitability. Effective human resource management helps minimize these risks. These internal factors are key causes of financial problems for many Russian entrepreneurs. To prevent such situations, it is important to develop competencies in financial management, strategic planning, and risk management.	Businesspeople may face financial problems for various reasons. Here are two of the most common. • **Insufficient planning and market analysis.** The absence of a detailed business plan that includes an analysis of the market, competitors, target audience, and financial forecasts can lead to poor decisions and inefficient use of resources. Unaccounted-for factors, such as changes in legislation, economic fluctuations, or the emergence of new competitors, can negatively impact the business's financial health. • **Ineffective financial management.** Problems with managing cash flows, the budget, and investments can lead to a lack of funds to cover expenses, debt to creditors, and other financial difficulties. The inability to allocate resources properly, optimize costs, and control finances can cause serious financial problems.

Table A2

The explanations in response to the question

"How can the effectiveness of expense control be improved?"

GigaChat 2.0	YandexGPT 5 Pro
The following methods can be used to improve the effectiveness of expense control. **Method 1: Budgeting** Description: Regular detailed budgeting allows one to plan expenses in advance and track actual costs against the	**Implementation of a budgeting and financial planning system.** • Development of a detailed budget for a specific period (month, quarter, or year) considering all expected revenues and expenses;

Table A2 (continued)

planned targets. This helps identify deviations and respond to them in a timely manner. **Advantages**: – sets spending limits for various categories; – ensures transparency of financial flows; – prevents overspending. **Method 2: Automating expense tracking and analysis** Description: Using special software and services to automate the expense tracking process simplifies the collection and processing of financial information; modern systems allow one to automatically classify transactions, generate reports, and analyze trends. **Benefits**: – reduces errors from manual data entry; – accelerates information processing; – enables rapid decision-making based on analytical reports. These methods help improve the accuracy and efficiency of expense control, ensuring the organization's financial stability.	• Regular supervision and comparison of factual expenses with budgeted values to identify variances and take corrective actions; • Use of software to automate the budgeting process and simplify the analysis of financial data. **Analysis and optimization of the expense structure:** • Regular analysis of all expense items to identify opportunities for optimization. This may include reviewing contracts with suppliers, seeking more favorable terms for purchasing goods and services, and eliminating unnecessary or inefficient expenses. Implementing lean manufacturing and cost management principles to reduce costs without compromising the quality of products or services. For example, processes can be optimized, losses decreased, and resource utilization improved.

Table A3

The explanations in response to the question

"How can budgeting be implemented in a small-sized business? List two methods."

GigaChat 2.0	YandexGPT 5 Pro
There are various methods for implementing budgeting in a small-sized business, the most common of which are the "top-down" and "bottom-up" methods. **The "top-down" method** With this method, company management defines overall goals and business development directions, sets key performance indicators (KPIs), and allocates resources among departments and employees. After top-level budgets are approved, departments develop their own detailed plans based on the set goals and allocated resources. Advantages of the method: • quick implementation, • clear allocation of responsibility, • the ability to consider management's strategic priorities. Disadvantages: • insufficient involvement of lower-level management staff, • potential lack of detail and realism in the budget. **The "bottom-up" method** This method assumes that department employees first form their own budgets based on current performance indicators, sales plans, costs, and other parameters. The data is then consolidated by management and approved at the top level of the company. Advantages of the method: • increased planning accuracy thanks to detailed knowledge of the situation in each department, • increased staff involvement in the budgeting process. Disadvantages: • a longer period for coordination and decision-making,	**Create a simple and clear budget with key revenue and expense items.** • Identify key revenue items (e.g., sales of goods or services) and expense items (rent, salaries, material purchases, utilities, etc.). • Use spreadsheets (such as Excel) to create the budget, as they simplify changes and data analysis. • Start by creating a short-term budget (such as a monthly or quarterly budget) to be able to adjust it quickly in response to changes in your business. **Appoint a person responsible for budgeting and budgetary supervision.** • Assign an employee or a team of employees to be responsible for preparing the budget, supervising it, and analyzing variances. • Ensure that the person in charge receives training in the basics of budgeting and financial analysis so they can effectively perform their duties. Hold regular meetings to discuss the current status of the budget, analyze variances, and plan corrective actions. This will help to control financial flows and respond to changes in a timely manner.

Table A3 (continued)

• the risk of conflicts and non-coordination between different departments. Thus, the choice of the appropriate method depends on the specifics of your small-sized business, staff readiness, and management requirements. It is recommended to combine both approaches—setting high-level targets at the top while simultaneously encouraging the participation of lower-level employees in developing specific plans.	

Table A4

The explanations in response to the question

"What is the cost of implementing budgeting in a small-sized business?"

GigaChat 2.0	YandexGPT 5 Pro
The cost of implementing a budgeting system in a small-sized business depends on several factors. Main costs 1. Software The choice of software plays a key role in determining the project cost. Here are some software options: • free solutions: there are free tools, such as Excel or OpenOffice Calc, but their functionality is limited and requires highly skilled staff; • entry-level paid products: for example, FinGrad, the Russian product starting at approximately 15,000–30,000 rubles, depending on functionality and the number of users; • special-purpose CRM/ERP systems with a budgeting module: these systems often include additional accounting and analytics modules, but prices start at 80,000–150,000 rubles, sometimes higher—around 200,000–300,000 rubles. Examples of Russian solutions: 1C Management of Our Firm, Bitrix24. 2. Consulting services Consultants, analysts, or auditors are often brought in to configure and adapt the software properly. The cost of services varies significantly: • The average hourly rate for an expert is around 2000–5000 rubles. • The project can take anywhere from a couple of weeks to a month or even longer, depending on the complexity of the business. • The total cost of consulting services will range from 100,000 to 500,000 rubles or more. 3. Staff training To ensure employees use the system effectively, it is important to conduct training: • in-house training by the company's experts: free of charge or at minimal cost; • external corporate training: typically costs between 10,000 and 50,000 rubles per group of participants. 4. Preparation of regulations and internal documents The process of preparing regulatory procedures also incurs costs: • If processes are developed in-house, this work is free. • Hiring external experts will cost additional 50,000–150,000 rubles. Additional factors affecting the price	The cost of implementing budgeting in a small-sized business can vary depending on several factors. • **Company size and the complexity of business processes:** the larger the scale of operations and the more complex the business structure are, the more resources will be required to implement budgeting. • **Availability of software:** using special budgeting software may require additional costs for purchasing licenses or developing a custom system. However, this can simplify the process and increase its efficiency. • **Staff training: t**he costs of training employees who will be involved in budgeting may also be included in the total cost. This may involve internal training or hiring external experts. • **Consulting services:** in some cases, companies may hire external consultants to assist with the development and implementation of the budgeting system, which will also add to the overall costs. • **The time and resources of management and employees:** implementing budgeting requires time and attention from management and key employees. These are internal resources to be considered as well. Overall, the exact cost figures depend on the particular conditions and needs of the business. For some small-sized businesses, implementing budgeting may incur minimal costs, especially if simple tools and existing resources are used; for others, the costs may be high due to the need to purchase special software and hire additional experts.

Table A4 (continued)

• Business size: the smaller the company is, the lower the requirements for software functionality and capabilities (and consequently, the cheaper the implementation) will be; • Automation of existing processes: the level of automation of accounting operations affects the complexity of budgeting integration; • Resource base: the availability of qualified personnel reduces the need to hire third-party experts. Overall cost estimate The minimum cost of implementing a basic budgeting project at a small-sized enterprise may amount to: • 70,000–100,000 rubles (self-configuration of free tools); • up to 500,000–700,000 rubles (for comprehensive implementation of special products, including consulting services and employee training). Thus, the final cost depends on the scope of work, the chosen software solution, and the availability of resources within the organization itself.	

Table A5

The explanations in response to the question
"Why do businesspeople encounter financial problems? Cite an economic law."

GigaChat 2.0	YandexGPT 5 Pro
Financial problems among entrepreneurs are often linked to a violation of the economic law of matching revenues to expenses. According to this law, financial well-being is ensured when revenues consistently exceed expenses, or when expenses are optimized to match the business's revenues. The main causes of financial problems **The mismatch between revenues and expenses:** • insufficient profitability of products or services; • high production costs or operating expenses; • low return on capital; • lack of financial discipline. **Incorrect risk assessment** Entrepreneurs sometimes underestimate risks such as changes in market conditions, a deteriorating economic situation, or the emergence of new competitors. This leads to financial losses and liquidity problems. **Capital management issues** Inadequate management of working capital can lead to a shortage of cash to cover current obligations. Effective management of inventory, accounts receivable, and accounts payable plays a crucial role. **Excessive debt burden** High interest rates on loans or suboptimal use of borrowed funds can significantly undermine the company's financial stability. The law of matching revenues and expenses: business operations are successful only when a balance is achieved between cash inflows and obligations to creditors, suppliers, and employees. A violation of this balance will inevitably lead to financial difficulties.	Businesspeople may face financial problems for various reasons related to both internal and external factors. One of the key economic laws explaining many of these problems is **the law of matching revenues and expenses**. **The main causes of financial problems:** 1. **The mismatch between revenues and expenses**: • When the company's expenses exceed its revenues, this inevitably leads to losses and financial difficulties. • Poor budget planning and a lack of control over expenses contribute to a cash shortfall. 2. **The insufficient profitability of products or services**: • If products or services are sufficiently profitable, this reduces the overall profitability of the business and can lead to losses. • Low profit margins on goods or services do not cover all costs and generate no profit. 3. **High production costs or operating expenses**: • Inefficient use of resources, high costs of raw materials, energy, rent, and salaries can significantly increase a company's expenses. • A lack of process optimization leads to rising costs. 4. **Low return on capital**: • Inefficient use of invested capital reduces the business's profitability and diminishes its ability to generate profit. 5. **Lack of financial discipline**: • Irregular financial accounting, lack of control over cash flows, and violation of the financial discipline lead to chaotic spending and financial problems. 6. **Incorrect risk assessment**: • Underestimating potential risks, such as changes in market conditions, competitors' actions, or economic crises, can lead to unforeseen losses.

Table A5 (continued)

	The economic Law: **The law of matching revenues to expenses** states that revenues shall consistently exceed expenses to ensure the company's financial stability and profitability. This means that: • A business must effectively manage its financial flows. • It is necessary to carefully plan the budget and control all expenses. • It is important to assess and minimize risks that could affect financial stability in a timely manner. Violation of this principle inevitably leads to financial difficulties, including a lack of funds to cover obligations, reduced liquidity, and, ultimately, the potential insolvency of the enterprise.

REFERENCES

1. Kulinich, A.A., Application of Large Language Models in Decision Support Systems. Part I: Explanation Models and Large Language Models, *Control Sciences*, 2026, no. 1, pp. 47–65.
2. *Inzhenernaya psikhologiya i ergonomika* (Engineering Psychology and Ergonomics), Klimov, E.A., Noskova, O.G., and Solntseva, N.G., Eds., 2nd ed., Moscow: Yurait, 2025. (In Russian.)
3. Hoffman, R., Mueller, S., Klein, G., and Litman, J., Metrics for Explainable AI: Challenges and Prospects, *arXiv:1812.04608*, 2018. DOI: https://doi.org/10.48550/arXiv.1812.04608
4. Marecek, D. and Rosa, R., Extracting Syntactic Trees from Transformer Encoder Self-Attentions, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, 2018, no. 1, pp. 347–349. URL: https://aclanthology.org/W18-5444/
5. Barskaya, I., Benchmarks for LLMs, *Unite.AI*, August 28, 2024. URL: https://www.unite.ai/benchmarks-for-llms/ (Accessed February 15, 2026.)
6. Serrat, O., The Five Whys Technique, in *Knowledge Solutions*, Singapore: Springer, 2017, pp. 307–310.
7. Bosselut, A., Le Bras, R., and Choi, Y., Dynamic Neuro-Symbolic Knowledge Graph Construction for Zero-Shot Commonsense Question Answering, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-21)*, Vancouver, Canada, 2021, vol. 35, no. 6, pp. 4923–4931.
8. Shwartz, V., West, P., Le Bras, R., et al., Unsupervised Commonsense Question Answering with Self-Talk, *arXiv:2004.05483*, 2020. DOI: https://doi.org/10.48550/arXiv.2004.05483
9. Likert, R., A Technique for the Measurement of Attitudes, *Arch. Psychol*., 1932, vol. 7, no. 140.
10. Tolstova, Yu.N., *Izmerenie v sotsiologii* (Measurement in Sociology), Moscow: Universitet, 2009. (In Russian.)
11. Goroshnikova, T.A., Computational Sociology: A Paradigm Shift or "Econometrics" of Sociology, *Humanities and Social Sciences. Bulletin of the Financial University*, 2021, vol. 11, no. 1, pp. 37–42. (In Russian.)
12. Mitina, O.V., Using the Logistic Equation Model to Study Dynamic Processes in Psychology and Life Sciences, *Modelling and Data Analysis*, 2013, no. 1, pp. 61–78. (In Russian.)
13. Thurstone, L.L. and Chave, E.J., *The Measurement of Attitude*, Chicago: The University of Chicago Press, 1929.
14. Cronbach, L.J., Coefficient Alpha and the Internal Structure of Tests, *Psychometrika*, 1951, vol. 16, no. 3. DOI: doi:10.1007/bf02310555.
15. *GigaChat-2*. URL: https://giga.chat/ (Accessed February 14, 2026; in Russian.)
16. *YandexGPT 5 Pro*. URL: https://ya.ru/ai/gpt (Accessed February 14, 2026; in Russian.)
17. Kitcher, P., Explanatory Unification and the Causal Structure of the World, in *Scientific Explanation*, Kitcher, P. and Salmon, W., Eds., Minneapolis: University of Minnesota Press, 1989, pp. 410–50.
18. Van Fraassen, B., *The Scientific Image,* Oxford: Oxford University Press, 1980.
19. Hempel, C., *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science,* New York: Free Press, 1965.

This paper was recommended for publication by RAS Academician D.A. Novikov, a member of the Editorial Board.

Received July 8, 2025, and revised November 17, 2025. Accepted December 16, 2025.

Author information

Kulinich, Aleksandr Alekseevich. Cand. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ alexkul@rambler.ru
ORCID ID: https://orcid.org/0000-0002-4751-205X

Cite this paper
Kulinich, A.A., Application of Large Language Models in Decision Support Systems. Part II: Measurement and Assessment of Decision-Maker's Satisfaction. *Control Sciences* **2,** 81–98 (2026).

Original Russian Text © Kulinich, A.A., 2026, published in *Problemy Upravleniya*, 2026, no. 2, pp. 93–113.

Translated into English by *Alexander Yu. Mazurov*,
Cand. Sci. (Phys.–Math.),
Trapeznikov Institute of Control Sciences,
Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com