Моделирование и анализ данных
2025, *15*(4), 7—26.
https://doi.org/10.17759/mda.2025150401
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2025, *15*(4), 7—26.
https://doi.org/10.17759/mda.2025150401
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

АНАЛИЗ ДАННЫХ | DATA ANALYSIS

Научная статья | Original paper

УДК 004.891:159.9.072.53

Evolutionary algorithms to generate prompts and verify responses of intelligent assistants

L.S. Kuravsky✉, D.A. Odintsov, M.A. Mikhailovsky
Moscow State University of Psychology and Education, Moscow, Russian Federation
✉ l.s.kuravsky@gmail.com

Abstract

Tools for working with intelligent assistants (in particular, with ChatGPT and DeepSeek) have been developed. A new mathematical framework, specialized algorithms, and software, which allow for validating the content of intelligent assistant responses and generating prompts based on annotations (brief descriptions of the responses to prompts), have been presented. These tools make it possible to replace a prompt engineer or, at least, automate his work. The solution in use is based on the evolutionary algorithms which generate a sequence of prompts organized according to a specific logical scheme and include a quasi-genetic algorithm with pseudo-crossover and pseudo-mutation operations, followed by analysis of the intelligent assistant's responses with the aid of multivariate statistical analysis methods. The search for an acceptable result, in which the intelligent assistant itself is actively involved, is an iterative process converging toward a given solution. The applied approach is justified and illustrated by examples of its use for solving psychological problems. The article is intended for programmers and mathematicians working with the large language models.

Keywords: intelligent assistant, large language model, artificial intelligence, evolutionary algorithm, quasi-genetic algorithm, prompt engineering, metric multidimensional scaling, psychology

For citation: Kuravsky, L.S., Odintsov, D.A., Mikhailovsky, M.A. (2025). Evolutionary algorithms to generate prompts and verify responses of intelligent assistants. *Modelling and Data Analysis, 15*(4), 7—26. (In Russ.). https://doi.org/10.17759/mda.2025150401

Introduction

Currently, the number of users of intelligent assistants (now commonly referred to as «artificial intelligence» or AI, or «neural networks») is growing exponentially. A significant portion of these users, having neither mathematical training nor programming experience, declare themselves to be advanced AI specialists. This creates problems due to attempts to misuse available intelligent tools, questionable interpretation and practical application of the results obtained, as well as misleading, grandiose statements without real content.

This makes it important to create user support tools for working with intelligent assistants (IA), and first and foremost, it is important to develop tools for automating the work of prompt engineers who prepare prompts for IA.

In general, the result of a subject's interaction with an IA is determined by two factors:

- the semantic content of prompts or other texts presented to the IA;
- the intellectual capabilities of the IA itself, which can vary widely.

The uncertainty of interpreting the semantic content of prompts or other relevant texts, as well as the known unpredictability of the AI's response to prompts, complicate the application of formalization, requiring its significant adaptation to the new context of application. This work is one of the first attempts at such adaptation. A special notation has been developed to ensure a compact description of the algorithms used.

The generation of plausible but incorrect information, known as «hallucination», remains a problem accompanying the practical use of AI. In particular, it is well known that AI can justify mutually contradictory statements if it receives a corresponding prompt. The only exceptions are strictly substantiated or obvious observable facts. Therefore, it is becoming urgent to find tools that can objectively evaluate the correctness of formulations calculated using AI.

The issues discussed in this paper have become particularly relevant after 2020 (Nikolenko et al., 2020), so there are relatively few relevant publications on the automation of AI and the elimination of the problem of «hallucinations». Among the approaches that inspire moderate optimism are dialogical methods, including the «Debate Game» (Irving et al., 2025) and the Chain-of-Verification Method (Shehzaad et al., 2025), which eliminates «hallucinations» by asking AI to reflect on its own answers and self-correct. However, these approaches are not based on a significant formalization and are hopelessly far from useful practical application.

This paper presents algorithms for solving two problems:

- generating prompts for which the annotations (brief descriptions of the answers to the prompt) are closest to the given description (the solution is provided by an evolutionary algorithm for selecting prompts);
- verifying the correctness of the intelligent assistant's answers (the solution is provided by an evolutionary algorithm for verifying the correctness of IA answers, or the «pendulum algorithm»).

The main components of the above algorithms that determine the calculated result are a newly developed quasi-genetic algorithm that ensures the expansion of the set of prompts, and a method of multidimensional metric scaling, a rigorous description of which is rarely found in publications. The quasi-genetic algorithm is constructed by analogy with the well-known genetic algorithm (Emelyanov et al., 2003), used to solve optimization problems and to train neural networks, with crossover and mutation operations replaced by pseudo-crossover and variation operations performed by AI, which are similar in context but fundamentally different in content.

The main principle implemented in the approach to solving problems is that the intelligent assistant performs all substantive operations related to extracting quantitative estimates from the material under study, followed by analysis of these estimates using methods of multidimensional statistical analysis, statistical hypothesis testing techniques, and other mathematical tools.

The tools, which operate based on the algorithms described below, are implemented in software based on the OpenAI API. These tools have been tested in pilot mode with psychological texts, demonstrating convincing results.

The most obvious prospects for practical application of the presented algorithms are in areas where concepts with significant variability in interpretation are used: in psychology, sociology, art history, and other humanities (Shoham et al., 2009; Nikolenko et al., 2020).

This article is intended for programmers who create tools for working with large language models and mathematicians who develop methods for the practical use of artificial intelligence capabilities.

Notation and basic concepts

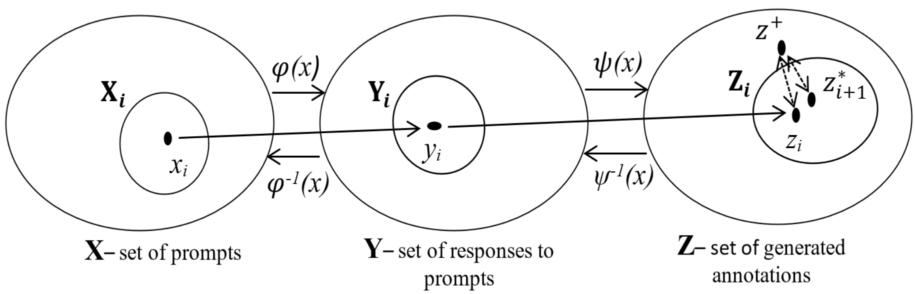

Fig. 1. Evolutionary prompt-selection algorithm for an intelligent assistant: operations on elements of metric spaces

X — set of prompts for the IA, **Y** — set of IA responses to prompts from set **X**, **Z** — set of IA-generated annotations from set **Y**. **X**, **Y**, **Z** — metric spaces with quasi-distances π_x, π_y, π_z (calculated for pairs of elements of spaces **X**, **Y**, **Z**, respectively, which are calculated on prompts for IA and represented by matrices $\chi(\ldots)$) and Euclidean distances ρ_x, ρ_y, ρ_z (calculated as a result of multidimensional scaling on matrices of pairwise quasi-distances and represented by matrices $\vartheta(\ldots)$ (see Fig. 1). Quasi-distances π_x, π_y, π_z may not satisfy the distance axioms (Kolmogorov, Fomin, 2023), while distances ρ_x, ρ_y, ρ_z satisfy these axioms. Thus, the metric of spaces **X**, **Y**, and **Z** is determined by Euclidean distances calculated because of multidimensional scaling of quasi-distances between non-numerical objects determined using AI. Multidimensional scaling provides filtering of inconsistencies in quasi-distance estimates and a transition from quasi-distances to mutually consistent distances (Borg, Groenen, 2005; Cox T, Cox M, 2001; Morrison, 1976; Rao, 1973).

Considering the capabilities of AI to generate prompts $x \in \mathbf{X}$ and AI mappings $\{\varphi(x)\}_{x\in\mathbf{X}}$ and $\{\psi(y)\}_{y\in\mathbf{Y}}$ to be virtually unlimited, the metric spaces **X**, **Y,** and **Z** can be considered complete.

X, **Y**, and **Z** are also used as scaling spaces. Distances are understood to be Euclidean distances. Quasi-distances are represented by values from the numerical interval [0;1] and are calculated for given pairs of elements of metric spaces **X**, **Y,** and **Z** using AI as the results of prompts for comparing elements included in given pairs. A value of 1 corresponds to a complete match between the contents of the compared elements, while a value of 0 corresponds to a complete mismatch (obviously, the comparison result is ambiguous and is determined by the characteristics of the AI used). AI prompts explicitly specify the requirement to compare the presented elements, expressing the result as a real number from the interval.

AI images $y = \varphi(x)$ and $z = \psi(y)$, where $x \in \mathbf{X}$, $y \in \mathbf{Y}$. The functions $\varphi(x)$ and $\psi(y)$ are implemented by AI on elements of spaces **X** and **Y**. The mappings $y = \varphi(x)$ and $z = \psi(y)$ provide the calculation, respectively, of the response to the prompt x and the annotation y using AI. $x = \varphi^{-1}(y)$ and $y = \psi^{-1}(z)$ — prototypes $z \in \mathbf{Z}$ and $y \in \mathbf{Y}$, which are defined only for already calculated images (the specified prototypes are saved for already calculated images).

- $z^+ \in \mathbf{Z}$ — specified annotation (brief description of the response to the prompt).
- $z^* \in \mathbf{Z}$ — current approximation to the annotation.
- Operations in metric space **X**:
- $\tau_{\mathbf{X}}(x)$ — the result of determining the neighborhood of prompt *x,* composed of elements of the set $\mathbf{X}$;
- $\eta(\mathbf{M})$ — the result of calculating the Kameni median for the specified set $\mathbf{M} \in \mathbf{X}$: $\eta(\mathbf{M}) = \underset{x\in\mathbf{M}}{\mathbf{argmin}} \sum_{r\in\mathbf{M}} \rho_x(x,r)$ (i.e., determining the average element of this set);
- $\chi(\mathbf{M})$ — the result of calculating the matrix $\|\pi_{ij} = \pi(m_i, m_j)\|$, where $m_j, m_k \in \mathbf{M}$, $i, j = 1,\ldots,n$, pairwise quasi-distances for n elements of the given set $\mathbf{M} \subseteq \mathbf{X}$ using IA while preserving the already formed mutual distances for previously marked elements **M**;
- $\chi^-(\mathbf{M})$ — the result of calculating the matrix $\|\pi_{ij} = \pi(m_i, m_j)\|$, where $m_j, m_k \in \mathbf{M}$, $i, j = 1,\ldots,n$, pairwise quasi-distances for n elements of a given set $\mathbf{M} \subseteq \mathbf{X}$ using IA without preserving the already formed mutual distances for previously marked elements **M**;
- $\mu_i(\boldsymbol{G})$ — the result of expanding the set $\boldsymbol{G}$ with the matrix of pairwise quasi-distances $\|\pi_{jk} = \pi(g_j, g_k)\|$, where $g_j, g_k \in \mathbf{G}$, by applying a quasi-genetic algorithm at its i-th iteration (i =1,2,…); the elements of the set $\mathbf{G}$ are analogous to “chromosomes” in classical genetic algorithms;
- $\vartheta(\mathbf{\Pi}(\mathbf{M}))$ — matrix of mutual distances $\|\rho_{ij} = \rho(g_i, g_j)\|$, where $g_j, g_k \in \mathbf{M}$, $i, j = 1,\ldots,n$, for n elements of a given set **M**, obtained as a result of multidimensional

scaling of the matrix of pairwise quasi-distances $\mathbf{\Pi} = \left\| \pi_{ij} = \pi\left(m_i, m_j\right) \right\|$, where $m_j, m_k \in \mathbf{M}$; multidimensional scaling provides filtering of inconsistencies in estimates of mutual quasi-distances π_x, π_y, π_z for pairs of elements of spaces **X**, **Y**, **Z** and transition from quasi-distances to their corresponding distances ρ_x, ρ_y, ρ_z; after this operation, the elements that have undergone multidimensional scaling are marked in order to fix the distances between them in the subsequent steps of the algorithm;

- $x_k = U\left(x_i, x_j\right) x_k = V\left(x_i, x_j\right), x_k = N\left(x_i\right)$ — pseudo-crossover operations performed using AI in space X ($x_i, x_j, x_k \in \mathbf{X}$), where U is a generalization of the content of two given texts, V is the selection of matching content from two given texts, N is the negation of the content of a given text;
- $x_k = \delta\left(x\right)$ — pseudo-mutation (variation) operation of a given text;
- $\mathbf{X}_i \subseteq \mathbf{X}$ — a subset of the space $\mathbf{X}$, calculated at the i-th iteration of the prompt generation algorithm;
- $\gamma\left(\mathbf{X}_i\right)$ — result of applying pseudo-crossover operations U, V and N, as well as pseudo-mutation δ to the set $\mathbf{X}_i$;
- $\omega(\mathbf{X}_i) = \mathbf{X}_i \cup \gamma\left(\mathbf{X}_i\right), \varphi\left(\omega(\mathbf{X}_i)\right), \ \psi\left(\varphi\left(\omega(\mathbf{X}_i)\right)\right),$ — extension of the set of prompts $\mathbf{X}_i$ by combining this set and supplementing it with $\gamma\left(\mathbf{X}_i\right)$, as well as the corresponding extensions of the sets $\mathbf{Y}_i = \varphi\left(\mathbf{X}_i\right)$ and $\mathbf{Z}_i = \psi\left(\mathbf{Y}_i\right)$;
- $\mathbf{X}_0$ — a set of basic prompts on a given topic;
- $i{+}{+}$ and $j{+}{+}$ — increase the indices i and j by one;
- ε — a small positive real number;
- $\alpha < 1$ — positive real number;
- I — generating prompt for the pendulum algorithm;
- t_i — content material (text or text with illustrations) being studied by the AI at the i-th iteration of the pendulum algorithm (i = 0, 1, 2, …);
- T_i^+ — a set of content materials consistent in content with t_0 at the i-th iteration of the pendulum algorithm (i = 0, 1, 2, …);
- T_i^- — a set of content materials that contradict the content of t_0 at the i-th iteration of the pendulum algorithm (i = 0, 1, 2, …);
- $\zeta\left(\mathbf{R}\right)$ — the centroid of objects, the mutual distances between which are represented by the mutual distance matrix $\mathbf{R}$ (the centroid coordinates are obtained by averaging the coordinates of the specified objects along each of the scaling space axes);
- $\Omega\left(\mathbf{R}\right)$ — dispersion of distances between objects, whose mutual distances are represented by the mutual distance matrix $\mathbf{R}$, to the centroid $\zeta\left(\mathbf{R}\right)$;
- N_{max} — natural number.

Evolutionary algorithm for prompt selection for AI

The task is formulated as follows.

Given: $\mathbf{X}_0$ — a set of basic prompts on a given topic; z^+ — a given annotation.

Find: $x^* = \mathbf{argmin}_x (\rho_z\left(g\left(f\left(x\right)\right), z^+\right)$

(i.e., find the prompt x^*, the annotation to the answer that is closest to the given description z^+).

Solution algorithm:

1. Set a set of basic prompts $\mathbf{X}_0$. $i=0$. $\mathbf{Y}_0=\{\varphi(x)\}_{\mathbf{x}\in\mathbf{X}_0}=\varphi(\mathbf{X}_0)$ and $\mathbf{Z}_0=\{\psi(y)\}_{\mathbf{y}\in\mathbf{Y}_0}=\psi(\mathbf{Y}_0)$ for all elements of the sets $\mathbf{X}_0$ and $\mathbf{Y}_0$. $z_0^*=\eta(\mathbf{Z}_0)$.
2. If $i>0$, then compute the IA-mappings $\mathbf{Y}_i=\{\varphi(x)\}_{\mathbf{x}\in\mathbf{X}_i}=\varphi(\mathbf{X}_i)$ and $\mathbf{Z}_i=\{\psi(y)\}_{\mathbf{y}\in\mathbf{Y}_i}=\psi(\mathbf{Y}_i)$ for all elements of the sets $\mathbf{X}_i$ and $\mathbf{Y}_i$.
3. Find the median of Kemeny $x_i=\eta(\mathbf{X}_i)$ and the images of its mappings $y_i=\varphi(x_i)$ and $z_i=\psi(y_i)$.
4. Compute the extension $\omega(\mathbf{X}_i)=\mu_i(\mathbf{X}_i)$, using *a quasi-genetic algorithm*.
5. Compute the matrices $(\chi(\omega(\mathbf{X}_i))$, $\chi(\varphi(\omega(\mathbf{X}_i)))$ and $\chi(\psi(\varphi(\omega(\mathbf{X}_i)))))$.
6. Compute $\vartheta(\chi(\omega(\mathbf{X}_i)))$, $\vartheta(\chi(\varphi(\omega(\mathbf{X}_i))))$, and $\vartheta(\chi(\psi(\varphi(\omega(\mathbf{X}_i)))))$, marking the resulting mutual distances between the elements of the sets $\mathbf{X}_i$, $\mathbf{Y}_i$, and $\mathbf{Z}_i$.
7. Find the Kameni median $\eta(\omega(\mathbf{X}_i))$.
8. Determine the neighborhood of the prompt $\tau_{\mathbf{X}_i}(\eta(\omega(\mathbf{X}_i)))$ using the mutual distances defined by the matrix $\vartheta(\chi(\omega(\mathbf{X}_i)))$.
9. Compute IA mappings $\mathbf{Y}_i=\varphi(\tau_{\mathbf{X}_i}(\eta(\omega(\mathbf{X}_i)))),\mathbf{Z}_i=\psi(\varphi(\tau_{\mathbf{X}_i}(\eta(\omega(\mathbf{X}_i)))))$ for all elements of the prompt neighborhood $\tau_{\mathbf{X}_i}(\eta(\omega(\mathbf{X}_i)))$.
10. Check the condition $(\exists z\in\mathbf{Z}_i)(\forall z_i\in\mathbf{Z}_i)((z_i\neq z)\,\&\,(\rho_z(z,z^+)<\alpha\rho_z(z_i,z^+)))$. If the condition is satisfied, then $z_{i+1}^*=\underset{z}{\mathbf{argmin}}\{\rho_z(z,z^+):\forall(z_i\in\mathbf{Z}_i)((z_i\neq z)\,\&\,(\rho_z(z,z^+)<\alpha\rho_z(z_i,z^+))\}$, delete the specified relative part of the elements $\mathbf{Z}_d=\{z_i\in\mathbf{Z}_i:\rho_z(z_i,z^+)>\rho_z(z_{i+1}^*,z^+)\}$ and their preimages $\mathbf{X}_{nd}=\varphi^{-1}(\mathbf{Y}_{nd})$ and $\mathbf{Y}_{nd}=\psi^{-1}(\mathbf{Z}_i\setminus\mathbf{Z}_d)$ and proceed to **step 11**, otherwise proceed to **step 4**.
11. Calculate $x_{i+1}^*=\varphi^{-1}(\psi^{-1}(z_{i+1}^*))$.
12. If $\rho_z(z_{i+1}^*,z^+)<\varepsilon$, then $x^*=x_{i+1}^*$ and **stop**, otherwise go to **step 13**.
13. $\mathbf{X}_{i+1}=\tau_{\mathbf{X}}(x_{i+1}^*)$.
14. i++.
15. Proceed to **step 2**.

Quasi-genetic algorithm for performing set expansion of prompts when performing operations $\mu_i(G)$

1. Set the set of basic prompts $\mathbf{X}_0$. *j=0.*
2. Check the condition $(\nexists x\in\mathbf{X}_j)(\rho_z(\psi(\varphi(x),z^+)<\rho_z(\psi(\varphi(\eta(\mathbf{X}_j))),z^+))$. If the condition is met, go to **step 3**, otherwise **stop**.
3. Quasi-genetic selection of elements of the set $\mathbf{X}_j$ at iteration j using the «roulette rule» with the distance $\rho_z(\psi(\varphi(x),z^+)$ as the quality function, where $x\in\mathbf{X}_j$.
4. Formation of the complement $\gamma(\mathbf{X}_j)$ to the set $\mathbf{X}_j$ by applying pseudo-crossover operations *U*, *V* and *N*, as well as pseudo-mutation δ , to the elements of the set $\mathbf{X}_j$.
5. Combining the set of prompts $\mathbf{X}_j$ corresponding to iteration j and the complement $\gamma(\mathbf{X}_j)$: $\omega(\mathbf{X}_j)=\mathbf{X}_j\cup\gamma(\mathbf{X}_j)$.
6. j++.
7. Go to **step 2**.

Consider the transformation $z^*_{i+1} = \Xi(z^*_i)$ defined by step 10 of the evolutionary algorithm above. According to the selection condition specified in the description of step 10, $\rho_z\left(z^*_{i+1}, z^+\right) < \alpha\rho_z\left(z^*_i, z^+\right)$. It is true that the sequence $\left\{z^*_j\right\}_j$ converges to z^+ in the space **Z**. Indeed, $\rho_z\left(z^*_j, z^+\right) < \alpha\rho_z\left(z^*_{j-1}, z^+\right) < \ldots < \alpha^j\rho_z\left(z^*_0, z^+\right)$. Since $\alpha < 1$, the value $\rho_z\left(z^*_j, z^+\right)$ is arbitrarily small for sufficiently large j at $\rho_z\left(z^*_0, z^+\right) > 0$ and equal to zero at $\rho_z\left(z^*_0, z^+\right) = 0$. Therefore, in the metric space **Z**, $z^+ = \lim_{j\to\infty} z^*_j$.

The presented version of the evolutionary algorithm demonstrated good convergence to the desired result. However, in a small neighborhood of the given description z^+, the distance $\rho_z\left(z^*_j, z^+\right)$ may converge not to zero, but to a small positive number, which is due to the limited variability of the results of the pseudo-crossover, given by the operations U, V and N. In this case, it is necessary to either expand the set of pseudo-crossover operations, increasing the variability of its results, or replace the stopping condition $\rho_z\left(z^*_{i+1}, z^+\right) < \varepsilon$ in step 12 of the evolutionary algorithm with a special case of the *Cauchy* condition for the convergence of numerical sequences: $\rho_z\left(z^*_{i+1}, z^*_i\right) < \varepsilon$.

The convergence of the computational procedure is determined by the condition specified in step 10 of its description, the result of which, in turn, depends on the semantic content of the prompts submitted by the AI and the intellectual capabilities of the AI itself.

Evolutionary algorithm for checking the correctness of AI responses

Main thesis: *AI responses that are correct in content are significantly more consistent with each other than responses that are incorrect in content.*

Avoiding philosophical discussions on the topic "What is truth?", we will assume that statements that are confirmed in practice by the results of observations are better consistent with each other than their negations. In other words, an indirect assessment of correctness is applied: it is assumed that formulations that are more consistent in content are more plausible than those that are less consistent. We will call this statement *the thesis of consistency*. The quantitative measure of consistency is the dispersion of IA images of responses in metric space. The significance of differences in dispersion is established by testing the standard null hypothesis using the F-test. The content material required for analysis t_0 is either obtained in response to a generating prompt I or is specified directly.

1. Obtain the content material necessary for analysis t_0 in response to the generating request I or ask t_0 directly. $i=0$. $T_0^+ = t_0$. $T_0^- = \varnothing$.
2. $i{+}{+}$.
3. Calculate $t_i = N\left(t_{i-1}\right)$.
4. If i is even, then $T_i^+ = T_{i-2}^+ \cup t_i$ and $T_i^- = T_{i-1}^-$, otherwise $T_i^- = T_{i-2}^- \cup t_i$ and $T_i^+ = T_{i-1}^+$.
5. Calculate $\vartheta(\chi^-(T_i^+ \cup T_i^-))$.
6. Calculate the variances $\Omega^+ = \Omega\left(\vartheta\left(\chi^-\left(T_i^+\right)\right)\right)$ and $\Omega^- = \Omega\left(\vartheta\left(\chi^-\left(T_i^-\right)\right)\right)$.

7. Assuming that the distributions of distances to centroids $\zeta^{+} = \zeta\left(\vartheta\left(\chi^{-}\left(T_i^{+}\right)\right)\right)$ and $\zeta^{-} = \zeta\left(\vartheta\left(\chi^{-}\left(T_i^{-}\right)\right)\right)$ are normal, test the null hypothesis of equality of variances Ω^{+} and Ω^{-} using the F-test for the statistic Ω^{+} / Ω^{-} if $\Omega^{-} > \Omega^{+}$, or for the statistic Ω^{+} / Ω^{-} if $\Omega^{-} < \Omega^{+}$. If the null hypothesis of the F-test on the equality of variances is not rejected and $i < N_{max}$, then proceed to **step 2**, otherwise proceed to **step 8**.
8. If the null hypothesis of equality of variances is rejected, then if $\Omega^{-} > \Omega^{+}$, conclude that the content material under investigation is correct t_0, or if $\Omega^{-} < \Omega^{+}$, conclude that the content material under investigation is incorrect t_0, otherwise consider the correctness of the specified material to be undetermined.

For sets of correct and incorrect content materials, Wilks' statistics and the associated F statistics (Borg, Groenen, 2005; Cox T, Cox M, 2001; Cramer, 1999; Morrison, 1976; Rao, 1973) is used to assess the degree of discrimination between them. If the specified discrimination is significant ($p < 0,05$ for the specified statistics), then the conclusion about the correctness of the content material under study t_0 is considered reliable; otherwise, additional analysis of the content material t_0 should be performed.

The convergence of the computational procedure under consideration is determined by the condition specified in step 7 of its description, the result of which, in turn, depends on the semantic content of the prompt generated to the AI (or the original content being studied) and the intellectual capabilities of the AI itself.

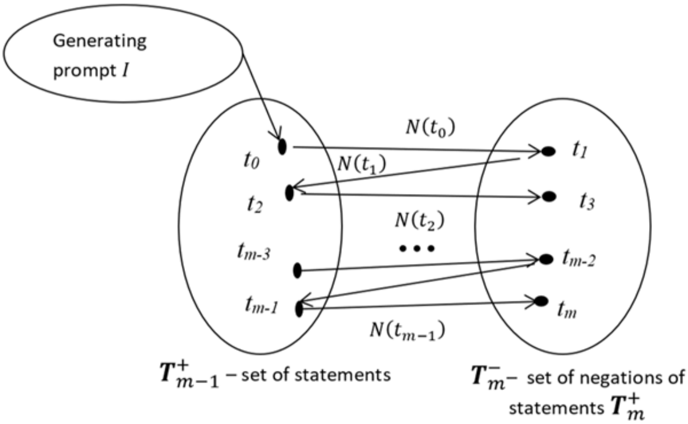

Fig. 2. Pendulum algorithm: operations on elements of the set of statements and the set of negations of statements

As part of the procedure for generating sets T_i^{+} and T_i^{-} in the scaling space, let us consider the set of events generating statements and their negations $\Gamma = \left\{\Gamma_i\right\}_{i=1}^{T_\Gamma}$. We will say that the subsets of events $O = \left\{O_i\right\}_{i=1}^{T_O} \in \Gamma$ and $\Phi = \left\{\Phi_i\right\}_{i=1}^{T_\Phi} \in \Gamma$ *are logically related* if the conditional probabilities $\left\{P(O_i \mid \Phi_1 \cup \ldots \cup \Phi_{T_\Phi})\right\}_{i=1}^{T_O}$ are sufficiently large, namely: $P\left(O_i | \Phi_1 \cup \ldots \cup \Phi_{T_\Phi}\right) \geq 1 - \delta$, where $\delta \ll 1$.

If we assume that the probabilities of autonomously considered events from the subsets O and Φ for all i satisfy the inequalities $P(O_i) \le \varepsilon$ and $P(\Phi_i) \le \varepsilon$, where $\varepsilon \ll 1$, then the probability of the sequential occurrence of the subsets Φ and O is equal to $\prod_{i=1}^{T_O} P(O_i | \Phi_1 \cup \ldots \cup \Phi_{T_\Phi}) \cdot \prod_{i=1}^{T_\Phi} P(\Phi_i)$ in the case of a regular connection and $\prod_{i=1}^{T_O} P(O_i) \cdot \prod_{i=1}^{T_\Phi} P(\Phi_i)$ in the case of its absence. The ratio of the second of these probabilities to the first, which can be called *the regularity index*, is a value of no less than the order of smallness than ε^{T_O}.

Thus, the following statement about the balance of probabilities is valid: *the probability of the appearance of subsets of events O and Φ in the case of a regular connection in ε^{-T_O} exceeds the probability of the appearance of the same subsets in the absence of such a connection.*

This statement can be interpreted as follows: when T_0 is sufficiently large, even a single manifestation of a certain regularity actually indicates its presence, and the value of T_0 can serve as a measure of the reliability of such a conclusion. For example, if T_0 is equal to 3 and ε is equal to 0.1, then the probability of a regular connection is approximately 10^3 times greater than the probability of its absence.

In particular, events can be considered as finding sets of points representing statements and their negations in certain areas of the scaling space or the distribution of certain configurations of points corresponding to the classes of features under study in specified areas of this space (Kuravsky, 2014; Kuravsky, Greshnikov et al., 2024; Kuravsky, Orishchenko et al., 2025; Kuravsky, Yuryev et al., 2024).

Torgerson's multidimensional metric scaling method

Since the description of this method in the above terms is not widely presented in publications, below is a description of one of its common variants — the classic method of multidimensional scaling by Torgerson (Borg, Groenen, 2025; Cox T., Cox M., 2001; Morrison, 1976). This method solves the problem of placing a set of points in a linear Euclidean space of a certain dimension, representing elements of a certain set **M**, according to a given matrix of pairwise quasi-distances between these elements.

The task, which generally does not have a unique solution, is to find, for a given quasi-distance matrix $\|\pi_{ij}(m_i, m_j)\|$, where $m_j, m_k \in \mathbf{M}, i, j = 1, \ldots, n$, for n elements of a given set **M**, their coordinates in a linear Euclidean space of dimension $m < n$, which corresponds to the Euclidean distance matrix $\|\rho_{ij}(m_i, m_j)\|$, where $m_j, m_k \in \mathbf{M}, i, j = 1, \ldots, n$, providing the minimum value of the criterion $S = \sum_{i=1}^{n}\sum_{j=1}^{n} \left|\pi_{ij} - \rho_{ij}\right|^2$.

Torgerson's multidimensional metric scaling algorithm

1. Calculate the matrix of quasi-distance squares $\mathbf{D} = \|d_{ij} = \pi_{ij}^2(m_i, m_j)\|$, where $m_j, m_k \in \mathbf{M},\ i, j = 1, \ldots, n$.
2. Calculate the matrix of mutual scalar products (*Gram matrix*) $\mathbf{B} = -\frac{1}{2}\mathbf{JDJ}$, where $\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{OO}^T$ is the double centering matrix, $\mathbf{I}$ is the unit matrix of size $n \times n$,

O is a column vector of n units, $\mathbf{O}^T$ is a row vector of n units (multiplication by the matrix $\mathbf{J}$ centers the matrix by subtracting the row mean and column mean from each of its elements and adding the overall mean).

3. Solve the algebraic eigenvalue problem by computing the spectral decomposition $\mathbf{B} = \mathbf{E}^T \mathbf{\Lambda} \mathbf{E}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2 \ldots,, \lambda_n)$ is a diagonal matrix of eigenvalues in descending order ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$), and $\mathbf{E}$ is a matrix of corresponding eigenvectors arranged in columns.
4. Compute the coordinate representation n of the elements of the given set $\mathbf{M}$ in a linear Euclidean space of dimension $m < n$ by defining the matrix $\mathbf{X} = =\mathbf{\Lambda}^{1/2}\mathbf{E}_m$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_m^{1/2})$ — diagonal matrix of square roots of m largest eigenvalues of the matrix $\mathbf{B}$, ordered in descending order, $\mathbf{E}_m$ — matrix of the first m coordinates of the eigenvectors of the matrix $\mathbf{B}$, arranged in columns (i.e., m first terms of the matrix $\mathbf{E}$; the coordinates n of the elements of the set $\mathbf{M}$ in the matrix $\mathbf{X}$ of size $m \times n$ are also arranged in columns).

It should be noted that:

- The Gram matrix $\mathbf{B}$ is symmetric and positive semidefinite, which allows it to be represented as $\mathbf{B} = \mathbf{X}^T\mathbf{X}$; one of the methods for calculating the coordinate matrix $\mathbf{X}$ is discussed above;
- It has been proven that the given algorithm provides a solution that ensures the smallest value of the criterion S;
- The value S is zero when $m = n - 1$;
- $\mathbf{Q}$ is an orthogonal matrix of size $m \times m$, and $\mathbf{X}$ is a certain k-coordinate solution to the problem that provides the smallest value of the criterion S. Then $\mathbf{QX}$ is an equivalent solution that provides the same smallest value of the criterion S.

Results

Generation of a prompt for which the annotation is closest to the given description

The prompts below retain the individual style of wording prepared by an experienced prompt engineer.

The text used as the given annotation z^+ was: «An *approach in psychology based on the study of observable behavior and its quantitative analysis through objective measurement methods*».

The set of $\mathbf{X}_0$ obtained because of a real dialogue with AI included the following basic prompts.

- What is the behavioral approach?
- How is behavior studied in psychology?
- What methods are used to analyze behavior?
- What is objective behavior study?

ChatGPT [gpt-4o-mini] was used to solve the problem.

As a result of an iterative process implementing an evolutionary algorithm, the following prompt was obtained: x^*. The annotation to the answer that is closest to the given description z^+ is: *«What are the advantages of behaviorism over other psychological schools?»*

The corresponding x^* AI response to the prompt ($\varphi\left(x^*\right)$ in the notation used) is the following text: «*Behaviorism is a psychological approach that focuses on the objective study of observable behavior and ignores internal psychological processes. It is distinguished by its use of experimental methods and its emphasis on the interaction of behavior with the environment*». Seven iterations of the procedure were performed. Figure 3 shows the dependence of the distance $\rho_z\left(z^*_{i+1}, z^+\right)$ on the number of iterations obtained in the calculation process.

To implement the pseudo-crossover operations U, V, N, pseudo-mutation δ, and quasi-distance estimation, the following prompts were sent to the AI.

Prompt *U*

«*Create a new prompt combining ideas from the following two prompts*».

Prompt *V*

«*Create a SHORT new prompt, highlighting the common key ideas from the following two prompts. The prompt should be concise (no more than 2—3 sentences)*».

Prompt *N*

«*Create a SHORT new prompt, supplementing the first prompt with elements from the second. The prompt should be concise*».

Prompt *δ*

«*Rephrase the following prompt BRIEFLY, preserving its meaning*».

Prompt *Π*

«*Assess the semantic distance between two texts on a scale from 0.00 to 1.00 as a measure of the difference between their main statements. Be strict: do not underestimate the distance, consider differences in details, level of abstraction, examples, and conclusions. If one text provides a general description and the other provides a specific theory or example, the distance should be increased, even if the topics are similar. Return only one number with two decimal places, without any additional text*».

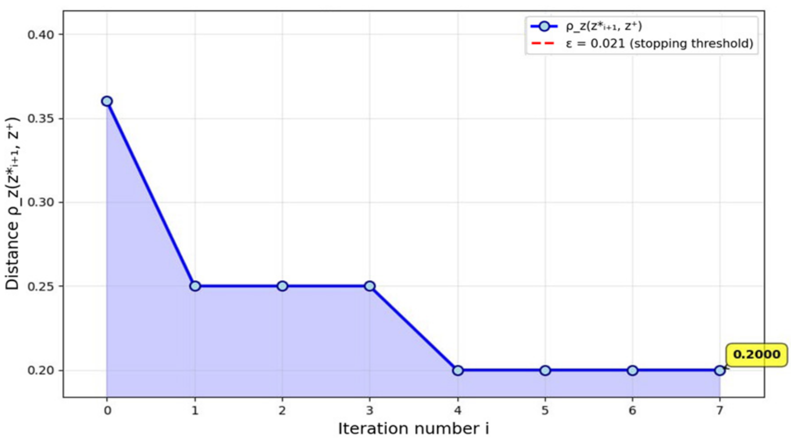

Fig. 3. Distance $\rho_z\left(z^*_{i+1}, z^+\right)$ as a function of the number of iterations

Verifying the correctness of a thesis using the pendulum algorithm

The prompts below retain the individual style of wording prepared by an experienced prompt engineer.

The phrase «*Cultural-historical psychology is based on Marxist theory*» was used as the thesis under investigation.

An excerpt from the *ChatGPT* communication log containing the prompts and the first elements of the sequence of meaningful responses $\{t_i\}_i$ is provided below.

Prompt 1 (generating prompt *I*)

«*You are an expert in philosophy, psychology, and scientific methodology. Formulate a clear, well-argued thesis H on a given topic that can be either confirmed or refuted; its plausibility will then be verified by the compactness of clusters of confirming and refuting arguments in the space of meanings. The answer should have a logical structure (thesis, justification, conclusion), be specific and concise (3—5 meaningful sentences)*».

Result 1

t_0 content material for analysis (received by IA)

«Cultural-historical psychology is based on Marxist theory.

Justification: The founder of cultural-historical psychology, Lev Vygotsky uses Marx's ideas about the social context in the development of the individual, emphasizing that psychological processes are formed as a result of social interaction and historical conditions. He emphasizes the role of culture and language in cognitive development, which is consistent with the Marxist concept of the influence of the social environment on the individual.

Conclusion: Thus, it can be argued that cultural-historical psychology is largely based on Marxist ideas, as confirmed by an analysis of the key concepts and principles underlying both approaches».

Request 2 $\left(\boldsymbol{N}\left(\boldsymbol{t}_0\right)\right)$ — refutation of substantive material t_0

«*You are working with a chain of theoretical statements.*

Formulate a new self-sufficient statement (1—2 sentences) that contradicts the meaning of the given text but remains within the original topic and subject area, preserving the key concepts and changing only the relationships between them.

Do not use metalanguage (such as "argument," "thesis," "this statement/text"), do not paraphrase or quote the original text, and do not offer compromises such as "on the one hand… on the other hand".

Return only the text of the new statement, without explanation».

Result 2

t_1: denial of substantive material t_0 (received by IA)

«Cultural-historical psychology cannot be reduced to Marxist ideas, as its basic principles, developed by Lev Vygotsky, emphasize the uniqueness of cultural factors and individual development, significantly differing from the Marxist emphasis on class struggle and economic conditions.

In fact, the understanding of cognitive development stems directly from the specific cultural context, rather than from Marxist theory about the influence of the social environment».

Request 3 $\left(\boldsymbol{N}\left(\boldsymbol{t}_1\right)\right)$ — refutation of content t_1 (confirmation of content t_0)
The text of request 3 coincides with the text of request 2.

Result 3
t_2 : confirmation of the content of t_0 (received by IA)
«Cultural-historical psychology is closely linked to Marxist ideas, as the emphasis on social conditions and interaction emphasizes that both individual development and cognitive processes cannot be understood without taking into account the economic and social factors described in Marxism. Thus, understanding these processes really depends on analyzing social structures and class relations, the importance of which significantly overlaps the role of unique cultural factors».

Request 4 ($N\left(t_2\right)$) — refutation of content t_2 (confirmation of content t_1)
The text of prompt 4 is identical to the text of prompt 2.

Result 4
t_3 : denial of content material t_0 (received by IA)
«Cultural-historical psychology can be understood outside of Marxist theory, as it focuses on unique cultural factors, forms of communication, and linguistic practices that shape cognitive processes independently of economic and social conditions. Thus, individual development and psychology can be studied without relying on class relations or social structures, which highlights the diversity of influences on personality and thinking styles».

Pairwise quasi-distances for affirmations and negations of content material t_0, forming the matrix $\chi^-\left(T_i^+ \bigcup T_i^-\right)$, were calculated using the following prompt.

«Estimate the semantic distance between two texts on a scale from 0.00 to 1.00 as a measure of the difference between their main statements. Be strict: do not underestimate the distance, consider differences in details, level of abstraction, examples, and conclusions. If one text provides a general description and the other provides a specific theory or example, the distance should be increased, even if the topics are similar. Return only one number with two decimal places, without any additional text».

The results of multidimensional scaling, obtained in the form of a matrix of mutual distances $\vartheta\left(\chi^-\left(T_i^+ \bigcup T_i^-\right)\right)$ at $i = 8$ and i =30, are presented as scatter plots in Figures 3 and 4. The complete separability of the sets T_i^+ and T_i^- in the scaling space confirms the correctness of the results obtained.

For a sample including 4 confirmations and 4 denials of the content material t_0 ($i = 8$), the F-test for statistics Ω^- / Ω^+ gives a value of 13.66 ($p < 0{,}03$), which, based on the accepted thesis of consistency, allows us to conclude that the content material under study t_0 is correct (i.e., cultural-historical psychology is indeed based on Marxist theory). For a sample including 15 confirmations and 15 denials of the content material t_0 ($i = 30$), the F-test for statistics Ω^- / Ω^+ gives a value of 6.17 ($p < 0{,}001$), which leads to the same conclusion.

At a significant level of $p = 0{,}05$ for testing the null hypothesis, the pendulum algorithm completes its work already at the 8th iteration. The result obtained is semantically correct.

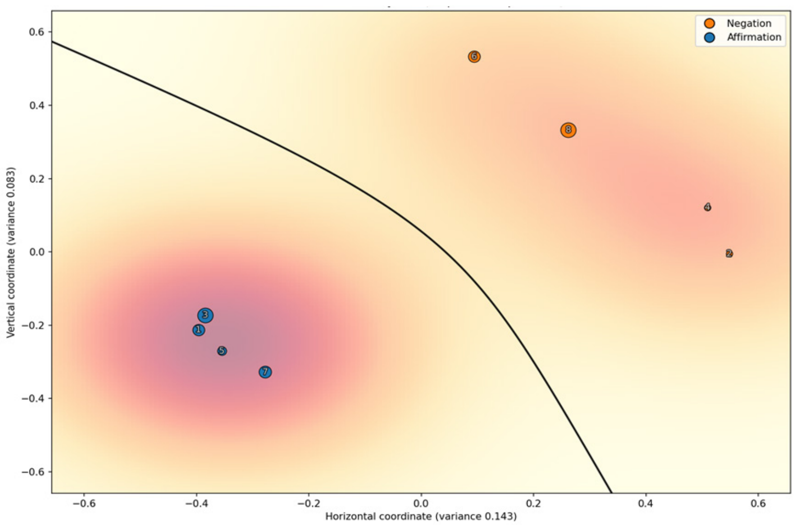

Fig. 4. Results of metric multidimensional scaling for $i = 8$

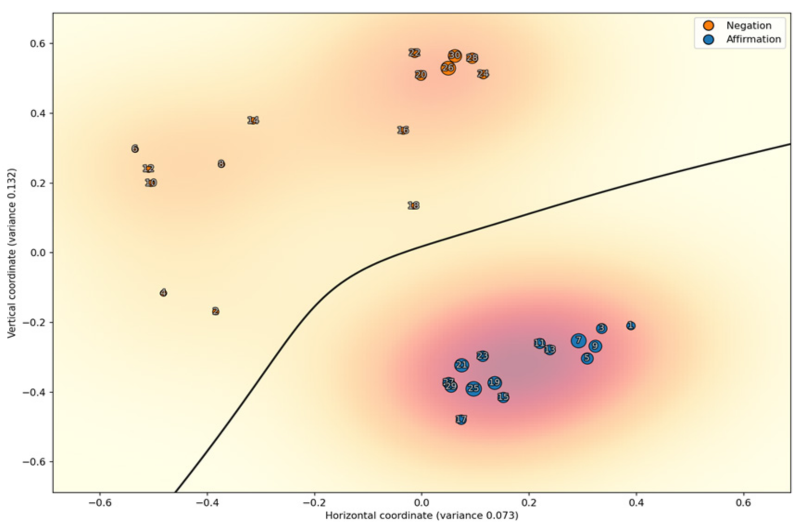

Fig. 5. Results of metric multidimensional scaling for $i = 30$

The duration of calculations on a medium-speed computer (processor base frequency — 2.70 GHz) was 71 seconds, with an average duration of 2.34 seconds per prompt to the AI.

Additional observation: narrowing of the chain $\{t_i\}$ to the semantic core of the contradiction

During the testing phase of the pendulum algorithm, an early version of the prompt specifying the operation $N(t_i)$ was used. In this formulation, the model received an instruction to formulate a «strict refutation» of the previous text without requirements to preserve the structure of the original thesis and without restrictions on the nature of the transformation of relations between key concepts, which led to an unforeseen phenomenon.

In the early stages (iterations 1—6), the AI's statements directly operate on the content of t_0, varying the arguments for and against the connection between cultural-historical psychology and Marxist theory.

However, starting from approximately the 15th-18th iteration, and particularly evident in the final iterations, the chain $\{t_i\}$ demonstrates a persistent narrowing of the semantic range. Instead of referring to a wide range of factors (historical materialism, the role of culture, the specificity of psychological mechanisms), statements begin to organize themselves around a single stable semantic opposition:

1. «Individual experience and unique cultural forms are considered a significant indicator of cognitive development».
2. «Individual experience is interpreted as a subjective basis, insufficient without analysis of social and structural conditions».

Thus, the operation $N(t_i)$, defined as the generation of a statement that contradicts the previous one, leads not only to an alternation of affirmations and negations, but also to the identification of a semantic attractor, which in this context is understood as a pair of statements that are an internal contradiction of the discourse assimilated by the AI.

This observation allows us to consider the pendulum algorithm as a tool that automatically identifies the semantic basis around which a dispute on a given topic takes place.

Conclusion

1. The result of interaction with the IA is determined by two factors: the semantic content of prompts or other texts presented to the IA, and the intellectual capabilities of the IA itself, which can vary widely.
2. The main principle implemented in the applied approach to solving problems is that the intelligent assistant performs all content-related operations associated with extracting quantitative estimates from the material under study, followed by analysis of these estimates using methods of multidimensional statistical analysis, statistical hypothesis testing techniques, and other mathematical tools.
3. An evolutionary algorithm has been developed for generating prompts for the IA, the annotations of which are closest to the given descriptions, as well as an evolutionary algorithm for verifying the correctness of the intelligent assistant's responses.

4. The basis of the evolutionary algorithm for generating prompts is a quasi-genetic algorithm, which ensures the expansion of the set of prompts. The quasi-genetic algorithm is constructed by analogy with the well-known genetic algorithm used to solve optimization problems and, in particular, for training neural networks, with the replacement of crossover and mutation operations with pseudo-crossover and variation operations performed by the IA, which are similar in context of application but fundamentally different in content.
5. The pendulum algorithm allows identifying the semantic basis around which the debate on a given topic takes place.
6. A special notation has been developed to ensure a compact description of evolutionary algorithms.
7. The convergence of the evolutionary algorithm for generating prompts under certain conditions (presented in step 10 of the algorithm description) has been proven, the result of which is determined by the semantic content of the prompts submitted by the AI and the intellectual capabilities of the AI itself.
8. The convergence of the evolutionary algorithm for checking the correctness of answers is determined by the condition specified in step 7 of its description, the result of which depends on the semantic content of the generating prompt to the AI (or the original content material being studied) and the intellectual capabilities of the AI itself.
9. It has been proven that with a sufficiently large number of events, even a single manifestation of a certain pattern actually indicates its existence. Such events can be considered as finding sets of points in certain areas of the scaling space.
10. The pilot application of the developed algorithms for solving psychological problems has demonstrated their effectiveness and semantic correctness.

References

1. Emelianov V.V., Kureichik V.V., Kureichik V.M. *Theory and practice of evolutionary research.* — M: FIZMATLIT, 2003, 432 p.
2. Kolmogorov A.N., Fomin S.V. *Elements of the theory of functions and functional analysis.* — M.: URSS, 2023. 572 p.
3. Kuravskiy L.S., Yuryev G.A., Mikhailovsky M.A., Nesimova A.O., Yuryeva N.E., Polyakov B.Yu. *Formation of teamwork skills and their objective quantitative assessment based on quantum representations* // Experimental Psychology. 2024. 17(2). P. 154—177. DOI: https://doi.org/10.17759/exppsy.2024170210.
4. Nikolenko S.I., Kadurin A.A., Arkhangelskaya E.O. *Deep learning*. — St. Petersburg: Piter, 2020. 480 p.
5. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling Theory and Applications* (Springer, New York, 2005).
6. Cox T.F. and Cox M.A.A., *Multidimensional Scaling*, 2nd ed. (Chapman and Hall/CRC, Boca Raton, 2001).
7. Cramer H., *Mathematical Methods of Statistics*. Princeton University Press, 1999. 575 pp.

8. Irving G., Christiano P., Amodei D. AI safety via debate. — *arXiv preprint, arXiv*:1805.00899. DOI: https://doi.org/10.48550/arXiv.1805.00899
9. Kuravsky L.S. *Quantum Representations and Their Applications in Diagnostics*. — Moscow: De Libri, 2024. 128 pp.
10. Kuravsky L.S., Greshnikov I.I., Kozyrev A.D., Kosachevsky S.G., Frolova L.I., Zakharcheva A.A. *A mathematical model for representing the related operator professional activities and its relevant diagnostic assessment based on the quantum representations*, Lobachevskii J. Math., 45 (6), 2534—2551 (2024).
11. Kuravsky L.S., Greshnikov I.I., Orishchenko V. A. *Quantum Representation of the Civil Aircraft Pilot Activity.* Lobachevskii Journal of Mathematics., 46 (6), pp.2609—2621, 2025.
12. Morrison D. F., *Multivariate Statistical Methods*, 2nd ed. (McGraw-Hill, New York, 1976).
13. Rao C.R., *Linear Statistical Inference and its Applications* (Wiley, Hoboken, 1973).
14. Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Raileanu Roberta, Xian Li, Asli Celikyilmaz, Weston Jason. *Chain-of-Verification Reduces Hallucination in Large Language Models*. — *arXiv preprint, arXiv*:2309.11495. DOI: https://doi.org/10.48550/arXiv.2309.11495
15. Shoham Y., Leyton-Brown K. Algorithmic, *Game-Theoretic, and Logical Foundations*. — London: Cambridge University Press, 2009.

Information about the authors

Lev. S. Kuravsky, Doctor of Science (Engineering), Professor, Dean of Computer Science Faculty, Moscow State University of Psychology and Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0000-0002-3375-8446, e-mail: l.s.kuravsky@gmail.com

Dmitry A. Odintsov, student, Computer Science Faculty, Moscow State University of Psychology and Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0009-0008-7082-700X, e-mail: dmitriyodintsov101@gmail.com

Michael A. Mikhailovsky, Research Assistant, Master's student, Computer Science Faculty, Moscow State University of Psychology and Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0000-0002-7399-2800, e-mail: muxa172002@yandex.ru

Contribution of the authors

All authors participated in the discussion of the results and approved the final text of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Поступила в редакцию 01.12.2025 | Received 2025.12.01
Поступила после рецензирования 03.12.2025 | Revised 2025.12.03
Принята к публикации 03.12.2025 | Accepted 2025.12.03
Опубликована 28.12.2025 | Published 2025.12.28