

HAAR – A Framework for Assessing Authorial Command of Scientific Reasoning Based on Semantic Networks

Dmytro Lande

ORCID: 0000-0003-3945-1178

National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” Kyiv, Ukraine

Abstract

The widespread adoption of large language models is substantially transforming the process of preparing academic and scientific texts and complicating traditional interpretations of authorship. Determining the origin of a text based on stylistic features does not address a more important question for academic assessment: whether the person submitting a work as their own actually demonstrates command of the system of scientific reasoning underlying it. This article proposes the HAAR (Human Academic Authorship Reasoning) framework, designed to generate evidence-based indicators of an author's command of the reasoning underlying an academic or scientific work.

The approach is based on the automated reconstruction of a semantic network of scientific reasoning, the identification of critical nodes and transitions – **milestones** – the generation of a personalized questionnaire containing unambiguous questions and plausible answer options, the conduct of a sequential adaptive assessment, and the aggregation of responses into an evidence graph. A mathematical model of HAAR is proposed that includes a typed reasoning graph, a transition criticality function, a question model, a delayed cross-check mechanism, and an aggregated consistency functional. The **Anti-Coaching** and **Delayed Cross-Check** principles are introduced to prevent the respondent from being progressively guided toward the expected reasoning during the assessment itself.

Unlike approaches aimed at detecting machine-generated text, HAAR does not produce an “authorship probability.” Instead, it generates a structured evidence profile characterizing the respondent's command of critical transitions in the scientific reasoning represented by the submitted work. Potential applications of the framework are considered for the defense of coursework, undergraduate and graduate qualification works, dissertations, as well as for the peer review of scientific publications.

A dedicated GitHub repository has been created to support reproducibility and further development of HAAR. It provides the HAAR Framework specification, methodological descriptions, experimental results, illustrations, and application examples: github.com/DmytroLande/HAAR

Keywords: HAAR, authorship, scientific reasoning, semantic network, large language models, academic integrity, adaptive assessment, knowledge graph, milestones, academic assessment.

1. Introduction

Generative models are already capable of producing academic texts whose quality and stylistic characteristics make them difficult to reliably distinguish from human-authored texts [1].

Experimental studies demonstrate the limited ability of experts to determine the origin of such texts [2]. The problem is further complicated by the fact that academic texts are increasingly produced through interaction between humans and language models [3]. Generative models can produce academic texts that are difficult to reliably distinguish from human-authored ones in terms of both quality and stylistic characteristics. Experimental studies indicate the limited capabilities of both automated tools and expert assessment in determining the origin of such texts. The problem is further complicated by the increasing prevalence of academic texts produced through collaboration between humans and large language models. Under these circumstances, binary classification of a text as human- or machine-generated cannot be regarded as a sufficient basis for assessing academic authorship.

For academic assessment, a more informative question is the extent to which a person submitting a work as their own demonstrates command of the system of scientific reasoning represented in that work. This involves not merely knowledge of the text's content, but the ability to reconstruct the reasoning that connects the formulation of the problem and initial assumptions to hypotheses, the selection of methods, the interpretation of obtained results, and the formulation of conclusions. Of particular importance is the ability to explain research decisions, justify choices among possible alternatives, and determine how changes in initial conditions would affect the obtained results.

On this basis, this paper proposes HAAR (Human Academic Authorship Reasoning), a framework in which a submitted academic or scientific work is transformed into a semantic reasoning network. Critical nodes and transitions are identified within this network and subsequently used to generate a personalized questionnaire designed to obtain structured evidence of the respondent's command of the reasoning underlying the submitted research.

This paper proposes HAAR (Human Academic Authorship Reasoning), a framework in which the submitted work is transformed into a semantic reasoning network, critical transitions are identified within this network, and a personalized questionnaire is generated on their basis [4,5].

The aim of this study is to develop a formal model and architecture of HAAR for obtaining structured evidence of an author's command of the scientific reasoning represented in a submitted academic or scientific work.

To achieve this aim, the following tasks are addressed:

- to develop a semantic model of scientific reasoning;
- to identify critical transitions and milestones;
- to propose a mechanism for generating unambiguous questions with answer options;
- to implement an adaptive questionnaire trajectory;
- to prevent the respondent from being coached during the assessment process;
- to develop a model for aggregating responses;
- to define a form of HAAR output that does not reduce authorship to a single automatically calculated indicator.

2. Materials and Methods

The mathematical model is introduced below as a formal representation of the assessment process. Rather than treating HAAR as a sequence of isolated procedures, the proposed model describes a set of successive transformations in which the representation of information changes at each stage. This makes it possible to trace the transition from the submitted scientific or academic text to a structured representation of its underlying reasoning and, subsequently, to evidence characterizing the respondent's command of that reasoning.

2.1. General HAAR Model

Let the input to the framework be a completed scientific or academic text

$$T = \{t_1, t_2, \dots, t_n\},$$

where t_i denotes structural or semantic fragments of the text containing statements, hypotheses, descriptions of methods, experimental results, interpretations, conclusions, and other elements of scientific reasoning. The purpose of HAAR is not to classify the text according to how it was produced, but to successively transform the textual representation of the research into a structure suitable for assessing the extent to which the respondent demonstrates command of the reasoning underlying the work.

The general HAAR model can be represented as a composition of successive operators:

$$T \xrightarrow{\Phi_G} G_R \xrightarrow{\Phi_M} M \xrightarrow{\Phi_Q} Q \xrightarrow{\Phi_I} A \xrightarrow{\Phi_E} G_E \xrightarrow{\Phi_A} R.$$

The first operator, Φ_G , reconstructs the scientific reasoning represented in the text and generates a typed semantic network G_R . Its nodes correspond to substantive elements of the research, including problems, assumptions, hypotheses, methods, data, experiments, observations, interpretations, results, and limitations, while its edges represent logical, methodological, evidential, and other meaningful relationships among these elements.

The operator Φ_M analyzes the structure of G_R and identifies the set

$$M = \{m_1, m_2, \dots, m_k\}$$

of critical elements and transitions (**milestones**) whose understanding is essential for reconstructing the logic of the research. This set may include, in particular, transitions from a problem to a hypothesis, from a hypothesis to a method, from an experimental result to its interpretation, as well as decision points involving alternative research choices.

Based on the set M , the operator Φ_Q generates a personalized questionnaire

$$Q = \{q_1, q_2, \dots, q_l\}.$$

Each question q_i is associated with a specific milestone or fragment of the semantic network. Thus, the questionnaire is determined by the structure of the particular work being analyzed rather than being constructed as a universal set of questions for a given subject area.

The operator Φ_I describes the process of interaction with the respondent. As the questions are presented sequentially, a set of responses is obtained:

$$A = \{a_1, a_2, \dots, a_l\}.$$

The order of questions may change depending on the responses already obtained. Thus, Φ_I implements not only the presentation of questions but also an adaptive questionnaire trajectory. Each response is considered together with information about which element of G_R it assesses and the type of task through which the response was elicited.

The operator Φ_E maps the accumulated responses back onto the structure of scientific reasoning and constructs the evidence graph G_E . At this stage, individual responses are transformed into localized evidence associated with critical transitions. Some transitions may receive consistent supporting evidence, whereas others may be characterized by insufficient evidence or repeated inconsistencies.

The final operator, Φ_A , aggregates the information represented in G_E and produces the resulting profile

$$R = \Phi_A(G, E).$$

The profile R characterizes the structure of the evidence obtained regarding the respondent's command of the scientific reasoning represented in the work. It is not an estimate of authorship probability and does not, by itself, provide an automatic decision regarding authorship of the text.

Thus, the overall logic of HAAR can be represented in a compact form as

$$\boxed{T \rightarrow G_R \rightarrow M \rightarrow Q \rightarrow A \rightarrow G_E \rightarrow R}$$

This sequence represents a transition from **the text as the final product of research** to **structured evidence of the respondent's command of the reasoning represented in that text**. It provides the basis for the subsequent construction of the semantic network, the procedure for selecting milestones, the generation of questions, and the aggregation of questionnaire results.

The overall architecture of the HAAR framework and the interaction between its principal components are presented in Figure 1. The diagram shows the transformation of a submitted scientific work into a semantic reasoning network, the identification of critical milestones, the

generation and validation of an adaptive questionnaire, and the subsequent transformation of the respondent’s answers into structured reasoning evidence used for human assessment [6].

Figure 1. Architecture of Interaction among the Main Components of the HAAR Framework
(HAAR $\rightarrow$ Evidence, Human $\rightarrow$ Decision)

The validated questions form an **Adaptive Questionnaire** that interacts sequentially with the **Respondent**. The obtained responses constitute **Answer Evidence** and are mapped back onto the corresponding elements and transitions of the reasoning network. Their consistency across different questions is represented by the **Reasoning Evidence Graph**, which distinguishes supported, uncertain, and contradictory reasoning transitions.

The accumulated evidence is processed by the **HAAR Assessment** component and presented to the **Human Assessor** as a structured profile of the respondent's command of the reasoning underlying the submitted work. The final academic or authorship-related judgment is made by the human assessor rather than by the automated system. Thus, the fundamental separation of responsibilities in HAAR can be expressed as

HAAR $\rightarrow$ Evidence, Human $\rightarrow$ Decision.

This architecture shifts the focus from identifying who physically produced the text to evaluating whether the respondent can consistently reconstruct, justify, modify, and defend the scientific reasoning represented in the submitted work.

2.2. Semantic Network of Scientific Reasoning

Unlike a conventional entity-based semantic graph, in HAAR the nodes represent elements of scientific reasoning.

$$G_R = (V, E, \tau_v, \tau_E, w),$$

where V is the set of nodes; $E \subseteq V \times V$ is the set of directed edges; τ_v is the node-typing function; τ_E is the edge-typing function; and w represents the weight characteristics of the network elements.

The node types are defined as

$$\tau_v(v) \in \{P, A, H, M, D, E_x, O, I, C, R, L, F\}.$$

Here, P denotes a problem; A – an assumption; H – a hypothesis; M – a method; D – data; E_x – an experiment; O – an observation; I – an interpretation; C – a claim; R – a result; L – a limitation; and F – a future research direction.

A typical fragment of the scientific reasoning trajectory can be represented as

$$P \rightarrow H \rightarrow M \rightarrow E_x \rightarrow O \rightarrow I \rightarrow R$$

The edges G_R are not necessarily causal in the strict mathematical sense. They may represent logical, evidential, methodological, semantic, or causal transitions between elements of scientific reasoning.

2.3. Hidden Transitions

Of particular interest in HAAR are transitions whose underlying logic is not fully explicated in the text. Suppose that the work explicitly presents a transition $v_i \rightarrow v_j$, whereas the actual research decision may have had the structure

$$v_i \rightarrow \{x_1, x_2, \dots, x_k\} \rightarrow v_j,$$

Where $x_1, x_2, \dots, x_k$ represent alternatives, selection criteria, rejected methods, or intermediate hypotheses. Let

$$H_{ij} = \{x_1, x_2, \dots, x_k\}$$

denote the **hidden transition set**. Such transitions are particularly useful for assessing authorial command of the work [7].

2.4. Identification of Milestones

Not all nodes and edges in the network are equally informative. For each potential milestone m_i , we introduce a criticality function:

$$K(m_i) = \alpha C_i + \beta D_i + \gamma N_i + \delta U_i,$$

where C_i denotes the structural centrality of the element; D_i – the dependence of the main results on that element; N_i – the non-triviality of the transition; U_i – the degree of uncertainty or the number of plausible alternatives; and $\alpha, \beta, \gamma, \delta$ are weighting coefficients [8].

The set of critical milestones is then selected as

$$M^* = \text{TopK}_{m_i \in M} K(m_i).$$

Priority is therefore given not merely to central terms in the work, but to critical transitions, such as

$$H_i \rightarrow M_j, \quad O_i \rightarrow I_j, \quad I_i \rightarrow R_j.$$

2.5. Questionnaire Generation

For each critical transition e_i , a question is generated in the form $q_i = (s_i, O_i)$, where s_i is the question stem and O_i is the set of answer options:

$$O_i = \{o_{i1}, o_{i2}, \dots, o_{ik}\}, \quad 4 \leq k \leq 6.$$

One option corresponds to the reconstructed reasoning of the work, while the remaining options should be semantically related and plausible alternatives. This distinguishes HAAR from conventional knowledge testing: preference is given to **WHY**, **WHY-NOT**, **ALTERNATIVE**, **COUNTERFACTUAL**, **REMOVE**, **CONTRADICT**, **LIMIT**, **TRANSFER**, and **EXTEND** questions.

2.6. Ensuring Question Unambiguity

One of the requirements for generating a HAAR questionnaire is that each question be unambiguous with respect to the element of scientific reasoning being assessed. A question should be associated with one specific node or transition in the semantic network and should not allow multiple equally valid interpretations. Formally, this requirement can be expressed as

$$|Target(q_i)| = 1,$$

Where $Target(q_i)$ denotes the set of semantic network elements whose knowledge or understanding is assessed by question q_i .

This requirement is particularly important for multiple-choice questions. The correct option must unambiguously correspond to the reasoning reconstructed from the text, whereas the alternative options should be substantively plausible but should not permit equally valid justification within the context of the analyzed work. This approach is consistent with general requirements for constructing multiple-choice assessment items, particularly with respect to the unambiguity of the correct answer and the quality of distractors [9].

Before a question is included in the questionnaire, it undergoes automated validation. If more than one justifiable answer option is identified, the assessment target is insufficiently defined, or formal cues make it possible to identify the correct answer without understanding the content of the work, the question is returned for reformulation:

$$q_i^{(0)} \xrightarrow{\text{validate}} \begin{cases} q_i, & |Target(q_i)| = 1, \\ \text{Rewrite}(q_i^{(0)}), & \text{otherwise.} \end{cases}$$

Thus, the validation procedure constitutes a separate stage of questionnaire generation and precedes direct interaction with the respondent.

2.7. Adaptive Formation of the Questionnaire Trajectory

In HAAR, the questionnaire is considered not as a fixed sequence of independent test items, but as a controlled process of traversing critical elements of the semantic network. Let q_i denote the current question and a_i the respondent's answer. The selection of the next question can then be represented by the operator

$$q_{i+1} = \Phi_Q(G_R, M^*, A_i),$$

where G_R is the semantic network of scientific reasoning, M^* is the set of selected milestones, and

$$A_i = \{a_1, a_2, \dots, a_i\}$$

is the response history accumulated up to the current point. This principle is consistent with the general concept of adaptive testing, in which the selection of the next item depends on the results of previous interactions with the respondent [10].

In HAAR, however, adaptivity is not used to suggest the correct answer or gradually guide the respondent toward it. Instead, it is used to determine the subsequent assessment trajectory through the semantic network. Once an answer has been received, the system records it, associates it with the corresponding milestone, and proceeds to the next question without informing the respondent whether the previous answer was correct.

We define this requirement as the **Anti-Coaching Principle**. It states that information resulting from the evaluation of one response should not become a cue for subsequent responses. Therefore, the evaluation result for a_i is used internally by HAAR to control the questionnaire trajectory but is not disclosed to the respondent until the corresponding assessment cycle has been completed.

2.8. Delayed Cross-Check of Responses

A separate issue arises when a respondent's answer contradicts the reconstructed reasoning of the work or does not provide sufficient evidence for assessing the corresponding transition. Immediately repeating the same question in a different formulation is undesirable because the new wording may effectively reveal the structure of the expected answer. In this case, the procedure would no longer assess the respondent's initial command of the reasoning, but rather their ability to use information obtained during the assessment itself.

To reduce this effect, HAAR introduces a **Delayed Cross-Check** mechanism. If an answer a_i associated with milestone m_j requires additional assessment, the system does not return to m_j in the next question. Instead, the assessment continues along other branches of the semantic network, and the repeated assessment is performed after a certain interval:

$$q_i(m_j) \rightarrow q_{i+1}(m_k) \rightarrow \dots \rightarrow q_{i+d}(m_j), \quad j \neq k, \quad d > 1.$$

The question $q_{i+d}(m_j)$ should not be a simple rephrasing of $q_i(m_j)$. Instead, it is generated using a different type of operation on the same network fragment. For example, the initial question may concern the rationale for selecting a particular method, whereas the delayed question may address the consequences of replacing that method with an alternative or changing one of the initial assumptions.

This mechanism makes it possible to obtain several relatively independent pieces of evidence concerning the same critical transition. At the same time, a single incorrect response is not regarded as sufficient grounds for a negative conclusion. More important is whether the inconsistency recurs when the same element of reasoning is assessed in a different context.

2.9. Formation of the Evidence Graph

The questionnaire results are aggregated not merely as the total number of correct and incorrect responses. Since each HAAR question is associated with a particular node or transition in the original semantic network G_R , the responses can be mapped back onto this network. On this basis, the **Reasoning Evidence Graph** is constructed:

$$G_E = (V, E^+, E^?, E^-),$$

where V corresponds to the relevant elements of the original reasoning network; E^+ contains transitions for which consistent positive evidence has been obtained; $E^?$ contains transitions for which the evidence is insufficient or contradictory; and E^- contains transitions for which repeated independent assessments have revealed persistent inconsistencies between the respondent's answers and the reasoning of the study.

To provide a quantitative representation of the evidence associated with milestone m_i , we introduce the local score

$$S(m_i) = \frac{\sum_{j=1}^{n_i} w_{ij} s(q_{ij})}{\sum_{j=1}^{n_i} w_{ij}},$$

where q_{ij} is the j -th question associated with milestone m_i ; $s(q_{ij})$ is the evaluation score for the corresponding response; w_{ij} is the question weight determined by its type and diagnostic significance; and n_i is the number of independent assessments of the corresponding milestone.

Since different milestones have different levels of importance within the structure of scientific reasoning, the local scores can be aggregated using the previously introduced criticality function $K(m_i)$. The overall consistency score is defined as

$$S_{\text{HAAR}} = \frac{\sum_{i=1}^N K(m_i) S(m_i)}{\sum_{i=1}^N K(m_i)}.$$

The value S_{HAAR} characterizes the degree of consistency between the respondent's answers and the structure of scientific reasoning reconstructed from the submitted work. However, this score **should not be interpreted as a probability of authorship**:

$$S_{HAAR} \neq P(\text{Authorship} \mid A, T).$$

For example, a value of $S_{HAAR}=0.91$ indicates a high degree of consistency between the obtained evidence and the critical transitions of the semantic network, but it does not imply a 91% probability that the respondent is the author of the work.

For this reason, the final output of HAAR is not a binary “author/non-author” classification, but a **structured evidence profile** that includes local milestone scores, the results of initial and delayed assessments, identified inconsistencies, and their locations within the semantic network. This profile is used by the expert as an evidence base for the final assessment of the respondent's authorial command of the scientific reasoning represented in the work.

3. Results

As a result of the study, the HAAR framework architecture was developed, integrating the reconstruction of the scientific reasoning structure, the identification of critical transitions, the generation of a personalized questionnaire, and the aggregation of responses into a structured evidence profile. Unlike approaches aimed at determining whether a text is of machine or human origin, HAAR characterizes the extent to which a respondent is able to reconstruct, justify, and modify the system of reasoning represented in a particular scientific work.

The proposed model implements the successive transformation of the initial text T into a semantic network G_R , a set of critical elements M^* , a personalized questionnaire Q , a set of responses A , an evidence graph G_E , and a resulting profile R :

$$T \rightarrow G_R \rightarrow M^* \rightarrow Q \rightarrow A \rightarrow G_E \rightarrow R.$$

Thus, the resulting framework is not a standalone authorship classification algorithm, but an interconnected procedure for moving from the structure of a particular work to structured evidence of the respondent's command of its scientific reasoning.

3.1. Personalization of the Assessment Procedure

One property of the developed architecture is that the assessment procedure depends on the content of the particular work being evaluated. Unlike conventional testing, the set of questions in HAAR is not defined in advance. Instead, it is generated after the semantic network of scientific reasoning has been reconstructed and its critical transitions have been identified.

For a text T , the personalized questionnaire can be represented as the result of a composition of two transformations:

$$Q(T) = \Phi_Q(\Phi_M(G_R(T))),$$

where $G_R(T)$ is the semantic network reconstructed from the text; Φ_M is the operator for identifying critical milestones; and Φ_Q is the operator for generating questions based on these milestones.

It follows that, for two substantively different scientific works T_1 and T_2 , different reasoning networks and, consequently, different sets of questions will generally be generated:

$$T_1 \neq T_2 \Rightarrow G_R(T_1) \neq G_R(T_2) \Rightarrow Q(T_1) \neq Q(T_2).$$

The last implication is not strict for all possible texts; however, it expresses the fundamental principle of HAAR: questions are determined not by the subject area in general, but by the reasoning structure of the particular work. Consequently, the questionnaire becomes individualized for each submitted text, reducing the possibility of preparing for the assessment through the mechanical memorization of a universal set of answers.

Another property of the proposed approach is the traceability of each question to an element of the source network. For every $q_i \in Q$, the mapping

$$\psi(q_i) \rightarrow m_j$$

is retained, where m_j is the milestone or critical transition for which the question was generated. This makes it possible to interpret not only the overall set of responses but also to localize positive, uncertain, and contradictory evidence directly within the structure of scientific reasoning.

3.2. Assessment of a Critical Transition in the Semantic Network

To demonstrate the operation of the proposed model, consider a generalized fragment of a semantic network that is independent of any particular subject area. Suppose that a hypothesis H_1 has been formulated in a study, method M_1 has been selected to test it, experiment E_1 has been conducted, observation O_1 has been obtained, and result R_1 has been formulated:

$$H_1 \rightarrow M_1 \rightarrow E_1 \rightarrow O_1 \rightarrow R_1.$$

Suppose also that an alternative methodological branch M_2 existed for the transition $H_1 \rightarrow M_1$, but was not implemented:

$$H_1 \rightarrow \begin{cases} M_1 \rightarrow E_1 \rightarrow O_1 \rightarrow R_1, \\ M_2 \rightarrow \dots \end{cases}$$

The source text may explicitly document only the selection of M_1 , while the criterion for rejecting M_2 remains partly implicit. In HAAR terminology, such a fragment constitutes a **hidden transition** and becomes a candidate for inclusion in the set of milestones.

For the initial assessment of this transition, a question is generated concerning the reason for selecting M_1 instead of M_2 . If the response is consistent with the reconstructed reasoning of the work, positive evidence is assigned to the transition $H_1 \rightarrow M_1$. However, a single such question is insufficient to establish the stability of the respondent's reasoning.

Therefore, the same transition can be assessed again by introducing a counterfactual modification of the initial conditions. For example, the respondent may be asked to determine how the research design should change if an assumption or constraint that motivated the selection of M_1 no longer holds. In this case, the assessment no longer concerns the reproduction of a formulation contained in the text, but rather the respondent's ability to transform a local fragment of the network

$$G_R \rightarrow G'_R$$

under changed conditions.

Thus, a single milestone can generate several functionally different assessments: reconstruction of the rationale, selection among alternatives, counterfactual modification, or transfer of a decision to changed conditions. Consistent results across these different forms of assessment provide stronger evidence of the respondent's command of the corresponding transition than a single correct answer.

3.3. Formation of the Evidence Profile

The results of individual assessments are aggregated at the level of the corresponding milestones and then mapped onto the evidence graph G_E . As a result, the HAAR output can be represented as a structured profile

$$R = \{E^+, E^?, E^-, C_f, A_r, D_c\},$$

where E^+ is the set of transitions supported by consistent positive evidence; $E^?$ denotes transitions for which the evidence obtained is insufficient or unstable; E^- represents transitions characterized by persistent inconsistencies; C_f contains the results of counterfactual assessments; A_r contains the results of alternative reasoning assessments; and D_c contains the results of delayed cross-checks.

Importantly, such a profile preserves the **structure** of the evidence obtained. Two respondents may achieve the same proportion of correct answers while exhibiting fundamentally different patterns of errors. In one case, uncertainty may concern peripheral elements of the work, whereas in another it may affect the transition from the main hypothesis to the selected method or from experimental results to the conclusion. The use of the evidence graph G_E makes it possible to capture and distinguish these cases.

For an aggregated representation of the results, several levels of consistency can be used. **Strong Consistency** corresponds to a situation in which the respondent consistently reconstructs most critical transitions, including under alternative and counterfactual assessments. **Consistency** characterizes the reconstruction of the main reasoning structure without systematic inconsistencies. **Partial Consistency** indicates that some significant transitions remain unstable or insufficiently supported. The **Insufficient Evidence** category is used when the available data are insufficient for a meaningful assessment, whereas **Significant Inconsistencies** indicates repeated inconsistencies across several critical transitions.

These categories are not classes for automatic authorship recognition. Rather, they characterize the degree and structure of consistency between the respondent's answers and the scientific reasoning reconstructed from the submitted work. The HAAR result should therefore be interpreted as an **evidence profile**, provided to the expert together with the evidence graph, local milestone assessments, and the history of cross-checks.

This constitutes a fundamental distinction between the proposed output and a numerical “authorship index”: HAAR localizes **which specific elements of scientific reasoning the respondent is able to justify and transform, which require additional assessment, and which exhibit persistent inconsistencies**. The final interpretation of this profile remains the responsibility of the human expert.

4. Discussion

The proposed HAAR framework changes the object of assessment compared with approaches aimed at determining whether an academic text is of human or machine origin. In AI-generated text detection tasks, the primary object of analysis is the text T itself, and the result is its assignment to one of the possible origin classes. In HAAR, the text is considered primarily as a source for reconstructing the structure of scientific reasoning. Consequently, the object of subsequent assessment is not a set of linguistic characteristics of the text, but the respondent's ability to reconstruct and justify the critical transitions within the reconstructed semantic network.

This distinction can be formally represented as a transition from the classification task

$$T \rightarrow \{\text{Human}, \text{AI}\}$$

to the sequence

$$T \rightarrow G_R \rightarrow M^* \rightarrow Q \rightarrow A \rightarrow G_E,$$

where the initial text T is transformed into a semantic network of scientific reasoning G_R , from which critical milestones M^* are identified; a personalized questionnaire Q is then generated, and the respondent's answers A are mapped onto the evidence graph G_E .

This formulation does not negate the value of methods for determining text origin, but shifts the emphasis from the question of how the text was directly produced to the question of command of the

scientific reasoning represented in the work. This distinction is particularly important for texts produced with varying degrees of involvement of generative models, since the mere fact of LLM involvement does not, by itself, determine the degree of independence of the author's scientific reasoning.

HAAR shares certain features with the traditional oral defense of an academic or scientific work, since in both cases the respondent is expected to answer questions about the content and justify the decisions made. At the same time, the proposed approach formalizes this process: each question is associated with a specific milestone, transition, or local fragment of G_R , and the resulting answer can be incorporated into the evidence graph. Consequently, the assessment procedure becomes traceable from an individual question to the particular element of scientific reasoning that it is intended to assess. Such formalization may complement oral defense practices, which are regarded as one means of assessing understanding, argumentation, and authorial command of a submitted work [11].

At the same time, HAAR results cannot be interpreted as direct proof of the historical fact that a particular person created the text. A respondent who did not author the work but has studied it in detail and developed a deep understanding of the methods used and the reasoning underlying the results may potentially demonstrate a high degree of consistency with the reconstructed network. Conversely, the actual author may provide individual unsuccessful answers because of the time elapsed since the research was conducted, the complexity of a particular question, or characteristics of the assessment setting.

Therefore, the output of HAAR should be interpreted as **evidence of authorial command** – structured evidence of the respondent's command of the scientific reasoning represented in the work. In this context, a distinction should be made between **text authorship**, **reasoning ownership**, and **scientific responsibility**. HAAR is directly focused on the second of these aspects and can provide only additional evidence for a broader assessment of authorship and scientific responsibility.

4.1. Application of HAAR in Academic and Scientific Assessment

One potential application of HAAR is the defense of academic and qualification works. Once a completed text has been submitted, the system reconstructs G_R , identifies critical milestones, and generates a personalized set of questions. An instructor or a member of an assessment committee can review the reasoning map, evaluate the generated questions, and use all or part of them during a written or oral defense. The responses obtained are used to construct G_E , which serves as an additional source of information for the expert but does not replace the expert's final decision.

The level and nature of the questions may depend on the type of work and the expected degree of research independence. For coursework and bachelor's theses, greater emphasis may be placed on **WHY** and **WHY-NOT** questions aimed at assessing understanding of major decisions and alternatives. For master's theses, the proportion of **COUNTERFACTUAL** and **TRANSFER** questions may be increased, requiring the respondent to modify the original reasoning structure or transfer the applied approach to changed conditions. In dissertation assessment, **REFUTE**, **ALTERNATIVE**, **REPRODUCE**, and **EXTEND** questions become particularly relevant because they make it possible to assess the respondent's ability to critically evaluate their own results, consider alternative research trajectories, reproduce obtained results, and identify directions for

further development of the research. This approach is consistent with the broader shift toward AI-resilient and process-based assessment, in which not only the final text but also understanding, argumentation, and the ability to defend research decisions are assessed [12].

Importantly, personalization of the questionnaire does not imply full automation of the assessment process. HAAR provides an information basis for assessment, while the instructor, reviewer, or assessment committee retains the ability to determine the level of difficulty, select the most significant milestones, ask additional questions, and interpret the resulting profile with regard to the subject area and the academic level of the work.

4.2. Limitations and Directions for Further Research

The proposed approach has several limitations that should be taken into account in its practical application. First, the quality of the generated questionnaire directly depends on the accuracy of the reconstruction of the semantic network G_R . Incorrect identification of hypotheses, methods, results, or relationships among them may lead to the selection of irrelevant milestones and, consequently, to questions that do not reflect the actual reasoning structure of the research. Therefore, the reconstruction of G_R itself requires validation, particularly for complex interdisciplinary works.

A second limitation is the need to calibrate question difficulty. The same type of transition within the semantic network may require different levels of argumentation depending on whether coursework, a bachelor's thesis, a master's thesis, or a dissertation is being assessed. Accordingly, question-generation parameters should take into account not only the structure of G_R , but also the level of academic assessment.

Another issue is the influence of factors that are not directly related to command of scientific reasoning. Language proficiency, stress during a defense, time constraints, or individual characteristics of oral communication may affect the form of a response. Therefore, a single unsuccessful answer should not automatically be interpreted as evidence of a lack of understanding of the corresponding transition. For this reason, HAAR incorporates the accumulation of multiple evidence items and the **Delayed Cross-Check** mechanism.

Another significant limitation is the possibility that a respondent may deliberately prepare for the assessment using an LLM. If a person has access to the work and models possible questions in advance, this may potentially improve their HAAR performance. At the same time, questionnaire personalization, the use of hidden transitions, counterfactual tasks, and delayed cross-checks make it more difficult to reduce such preparation to the memorization of predefined answers. The robustness of HAAR against such strategies requires separate experimental investigation.

For empirical validation of the framework, it would be appropriate to form at least three groups of participants: the actual authors of the works G_1 , knowledgeable non-authors G_2 who have specifically studied the corresponding works, and independent readers G_3 :

$G_1 = \text{Authors,}$ $G_2 = \text{Knowledgeable Non – Authors,}$ $G_3 = \text{Independent Readers.}$

Such an experimental comparison would make it possible to investigate an important question for HAAR: whether these groups differ not only in the proportion of correct answers but also in the **structure and dynamics of their trajectories through the semantic network**.

Accordingly, individual profile components may be analyzed as dependent variables: S_{WHY} , representing the ability to justify research decisions; S_{WHYNOT} , representing understanding of rejected alternatives; S_{CF} , representing performance on counterfactual transformations; $S_{TRANSFER}$, representing the ability to transfer reasoning to changed conditions; and S_{DELAY} , representing the stability of responses under delayed cross-checking. In addition, the number and localization of E^+ , $E^?$, and E^- transitions may be analyzed.

The purpose of such an experiment would not be to establish a predetermined distinction between authors and non-authors, but to **empirically evaluate the discriminative properties of HAAR**, calibrate the milestone criticality function, determine the informativeness of different question types, and assess the robustness of the proposed procedure. Until such a study has been conducted, HAAR should be regarded as a formally defined method for generating evidence of command of scientific reasoning rather than as a validated tool for automatically determining authorship.

5. Conclusions

This paper proposes the HAAR (Human Academic Authorship Reasoning) framework for generating structured evidence of a person's command of the scientific reasoning represented in a submitted academic or scientific work. Unlike approaches primarily aimed at determining whether a text is of human or machine origin, HAAR shifts the object of assessment from the characteristics of the text itself to the underlying reasoning structure and the respondent's ability to reconstruct, justify, and transform that structure.

The developed formal model of HAAR describes the successive transformation of the initial text T into a semantic network of scientific reasoning G_R , a set of critical milestones M^* , a personalized questionnaire Q , a set of respondent answers A , an evidence graph G_E , and a resulting profile R :

$$T \rightarrow G_R \rightarrow M^* \rightarrow Q \rightarrow A \rightarrow G_E \rightarrow R.$$

The central object of the model is the typed semantic network G_R , whose nodes and relationships represent research problems, assumptions, hypotheses, methods, experiments, observations, interpretations, results, and limitations. Critical milestones and hidden transitions are identified within this network, including those associated with method selection, alternative research decisions, and the conditions governing transitions between individual elements of reasoning. These elements provide the basis for generating a personalized questionnaire whose questions can be traced to specific fragments of the semantic network.

The proposed **Anti-Coaching** and **Delayed Cross-Check** principles separate assessment from learning during the questionnaire process. A critical transition is reassessed not by immediately rephrasing the preceding question, but after other branches of the network have been traversed and by using a different type of task. This makes it possible to accumulate relatively independent

evidence regarding the same element of scientific reasoning and reduces the influence of cues that may arise from the assessment procedure itself.

The results of individual assessments are aggregated into a **Reasoning Evidence Graph**, in which transitions with consistent positive, insufficient, or contradictory evidence are localized. This representation makes it possible not only to characterize the overall consistency of the responses, but also to determine which specific fragments of scientific reasoning the respondent is able to justify, modify, and defend, and which require further assessment.

Importantly, the resulting profile is not interpreted as a probability of authorship and is not used to make an automatic binary “author/non-author” decision. HAAR provides an evidence base for expert assessment, while the final decision remains with an instructor, reviewer, members of an assessment committee, or other authorized experts. Thus, the conceptual principle of the framework can be formulated as **HAAR provides evidence; humans make the decision**.

To support reproducibility and practical use, a dedicated HAAR GitHub repository has been created: github.com/DmytroLande/HAAR. It provides the framework specification, methodological documentation, experimental results, illustrations, and application examples, serving as an open resource for reproducing experiments and further developing HAAR.

The proposed approach may be applied to the defense of coursework, bachelor's and master's theses, dissertations, and as an additional instrument for the assessment of scientific publications. At the present stage, however, HAAR should be regarded as a formalized framework for generating evidence of command of scientific reasoning whose discriminative properties require experimental validation.

Future research should focus on comparing HAAR performance among actual authors, knowledgeable non-authors, and independent readers; evaluating the informativeness of different question types; determining a sufficient number of milestones; calibrating the weights of the milestone criticality function $\backslash(K(m_i)\backslash)$; investigating the robustness of the **Delayed Cross-Check** mechanism; and analyzing the effects of prior respondent preparation using LLMs. Such experimental validation should establish the applicability limits of HAAR and the conditions under which it can be used in real-world academic and scientific assessment procedures.

References

1. Hadra, M., Cambridge, K. and Mesbah, M., 2026. Evaluating the accuracy and reliability of AI content detectors in academic contexts. *International Journal for Educational Integrity*, 22(1), p.4.
2. Yeadon, W., Agra, E., Inyang, O.O., Mackay, P. and Mizouri, A., 2024. Evaluating AI and human authorship quality in academic writing through physics essays. *European Journal of Physics*, 45(5), p.055703.
3. Richburg, A., Bao, C. and Carpuat, M., 2024, May. Automatic authorship analysis in human-AI collaborative writing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1845-1855).

4. Tsigaris, P. and da Silva, J.A.T., 2026. AI detecting AI in academic writing: Why most AI detector findings are false. *Next Research*, p.101396.
5. Amnie, A.G., 2026. AI-resistant and AI-resilient assessment in higher education: a systematic review of validity-grounded strategies, institutional frameworks, and equity implications. *Assessment & Evaluation in Higher Education*, pp.1-23.
6. Al Khatib, K., Ghosal, T., Hou, Y., De Waard, A. and Freitag, D., 2021, June. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing* (pp. 56-65).
7. Koslowski, B., 2012. Scientific reasoning: Explanation, confirmation bias, and scientific practice. *Handbook of the psychology of science*, 151, p.192.
8. Zhao, Q., Liu, J., Sullivan, N., Chang, K., Spina, J., Blasch, E. and Chen, G., 2021, April. Anomaly detection of unstructured big data via semantic analysis and dynamic knowledge graph construction. In *Signal processing, sensor/information fusion, and target recognition XXX* (Vol. 11756, pp. 126-142). SPIE.
9. Pham, H., Besanko, J. and Devitt, P., 2018. Examining the impact of specific types of item-writing flaws on student performance and psychometric properties of the multiple choice question. *MedEdPublish*, 7, p.225.
10. Zhuang, Y., Liu, Q., Bi, H., Huang, Z., Huang, W., Li, J., Yu, J., Liu, Z., Hu, Z., Hong, Y. and Pardos, Z.A., 2026. Survey of computerized adaptive testing: A machine learning perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
11. Elkhodr, M. and Gide, E., 2026, August. Assurance by design: embedding the SAGE Defend step in AI-integrated higher education assessment. In *Frontiers in Education* (Vol. 11, p. 1872630). Frontiers Media SA.
12. Kickbusch, S., Ashford-Rowe, K., Kemp, A., Boreland, J. and Huijser, H., 2025. Beyond detection: Redesigning authentic assessment in an AI-mediated world. *Education Sciences*, 15(11), p.1537.