

Labora. Roga. Proba.

A Level-Conditioned Oral Verification Protocol for AI-Assisted Homework

Author: Adrian STAN

ORCID: <https://orcid.org/0009-0003-1457-5155>

Date: 01.10.2026

Abstract

Labora. Roga. Proba. (LRP) is a protocol through which the teacher allows AI to be used for homework, at a **level of cognitive friction** set in advance, and then verifies understanding through random oral checks, with questions generated **according to that level**.

The protocol is theoretically grounded in the Mechanism of Cognitive Stimulation (MCS) and practically implemented as a **10-step slider** that determines which stages of the work belong to the student and which are performed by the AI. LRP reads this division of labor in reverse: each level indicates the stage that the teacher's question will later verify, during the oral check. The student's answer is then judged using a quality scale adapted from the SOLO taxonomy.

The proposed contribution is verification conditioned by **level**: the same text, with a different required level, produces different questions. The protocol requires no infrastructure, stores no student data, and leaves grading entirely to the teacher. The protocol does not claim to have been empirically validated; the paper proposes it as an open tool, under the CC BY 4.0 license, and also describes what should be measured to validate it.

1. The Problem

Homework done at home no longer proves that the student knows anything. An AI model produces a correct text in a few seconds, and the teacher cannot distinguish, from the submitted text, the student's work from the machine's.

The effect on learning is documented. In a study of more than a thousand high school students, those who learned mathematics with an AI that gave direct solutions later performed worse, without AI, than the group that had not used AI at all ([Bastani et al., 2025](#)). Another study shows that users systematically overestimate the performance they achieve with AI ([Fernandes et al., 2026](#)), a phenomenon also described in the paper **False Cognitive Power Transfer FCPT** ([Stan, A., 2025](#)), while intensive use of AI tools correlates negatively with critical thinking, with a stronger effect among young people ([Gerlich, 2025](#)).

Institutional responses so far have taken three forms, all with known limitations:

1. **AI-text detectors** err in both directions and are outpaced by each new generation of models.
2. **Bans** push AI use underground without eliminating it.
3. **Oral exams** verify understanding directly but do not scale: a 25-minute oral exam for a class of 36 students, assessed by two examiners, requires roughly 30 hours of work ([Ipeirotis and Rizakos, 2026](#)).

LRP starts from the premise that **AI will be used anyway**. The question is no longer whether the student uses AI, but how they use it and whether, in the end, they understand what they submitted as homework.

2. Context: The Mechanism of Cognitive Stimulation (MCS)

MCS determines **how much** of the cognitive work of a task stays with the human and how much the AI takes over. It is proposed in the [open letter \(cognitiveslider.org\)](https://cognitiveslider.org) as a **second parameter** of language models, alongside temperature, and is currently available as:

- a) a **browser extension with a visible slider**;
- b) a **set of instructions** in Markdown, copied into the AI assistant's settings ([implementation page](#)).

MCS **applies to the relationship between human and AI**, not to the model's capability. The level does not change how correctly the AI answers, but **who performs each stage** of the work: setting the goal, choosing the method, execution, and verification. Level 00 means the mechanism is off. Steps 01-10 progressively shift the stages from the AI to the human.

Level	Goal	Method	Execution	Verification
00	AI	AI	AI	AI
03	student	AI	AI	AI
05	student	student chooses from options	AI, step by step	together
07	student	student	student attempts, AI gives hints	AI
10	student	student	student	student; AI only asks and flags

The table shows a few representative steps; the full definition of all 10 steps is in the MCS v1.1 instructions and is repeated in Section 4. At high levels, the AI never refuses to answer: on explicit request it provides the answer (at level 10, after asking for a single attempt), and the user can lower the level at any time.

3. The Labora. Roga. Proba. Protocol

LRP has **three stages**, which also give the protocol its name: the student works (*labora*), the teacher asks (*roga*), the student proves understanding (*proba*). The target segment is upper secondary education (ISCED 3, roughly ages 14-18).

Workflow:

1. **The teacher sets the MCS level** for the assignment, according to its purpose. If the assignment aims for the student to learn a method, the level is higher. If the method is merely a tool for something else, the level is lower.
2. **The student may do the homework with the AI assistant** configured at that level, via the **extension (slider)** or via the **instructions**, and submits it electronically.
3. **The teacher randomly selects a few students** in the next class. Random selection is essential: no student knows whether they will be called on.
4. **The teacher generates the question**: they fill in a form with the subject, year of study, assignment prompt, MCS level, task type, and an excerpt from the student's homework, and the teacher's AI formulates the question (Appendix A).
5. **The teacher asks the question and assigns the grade**. The AI does **not** take part in grading.

Design principles:

1. **The outcome is verified, not the process**. The protocol does not try to prove whether the student used the slider or which level they used; it verifies whether the student now has the competence that working at the required level builds.
2. **Deterrence comes from randomness**. As with traditional oral questioning, the chance of being checked motivates every student, without all of them being checked.

3. **The AI formulates, the human judges.** Any reference point generated by the AI is **support for the teacher**, not a verdict.
4. **The mediation is symmetric.** The student places an AI between themselves and the homework; the teacher places an AI between the homework and the assessment. Both are **configured from the same** MCS matrix.
5. **The protocol adds no intermediaries.** It has no server, accounts, or activity logs; data reaches only the AI assistants that the student and the teacher already use.

4. Level-Conditioned Questions

The same homework excerpt produces different questions depending on the required MCS level. This is LRP's central contribution: the MCS matrix, which states which stage belongs to the student, is read in reverse and states which stage must be verified.

Level	What the student had to do	What the question verifies
00	nothing; mechanism off	the protocol does not apply
01	received the answer and the principle	explaining the principle behind the solution
02	made the decisions flagged by the AI	one decision made and its effect on the result
03	set the goal or the constraint	the goal or constraint and its influence on the result
04	choose among 2-3 approaches	the existing alternatives and the reason for the choice
05	chose and justified the approach; checked the result together with the AI	justification of the approach and a spot check
06	proposed the steps after the first one	continuing the solution from a given point
07	attempted first, received hints	redoing a step under modified conditions
08	led each key step	the sequence of key steps and how they connect
09	proposed the method and the solution	applying the method to a variant of the problem
10	built the method, the solution, and the verification	verifying correctness and the points of risk

The teacher's AI does not know at which level the student actually worked, only the level required of them. That is why **the question does not measure the process, but the competence** that the process should have built. A student who worked at level 00 and then thoroughly understood the homework can pass the check; for the school, this outcome is acceptable.

A consequence of the construction: the questions at levels 07 and 09 require application in a modified situation, that is, transfer. High MCS levels build transfer, so their verification also tests it.

5. Assessment Instruments

Through the "task type" variable, the teacher chooses **how detailed** the reference received from the AI is. The three types draw on assessment forms already used internationally, from simple to complex.

Type	What the teacher receives	Assessment form adopted
A (quick)	the question only	the teacher's holistic judgment
B (standard)	the question, the stage verified, indicative content	indicative content, as in British mark schemes
C (complete)	as in B, plus levels of response and a second question	levels-of-response mark scheme, adapted from the SOLO taxonomy

Indicative content (type B). British exam mark schemes give the examiner ideas that a good answer may contain, state explicitly that the list is not exhaustive, and accept any other correct answer ([example Edexcel mark scheme](#)). LRP adopts this wording so that a correct but original answer is not penalized.

Levels of response (type C). The SOLO taxonomy describes the quality of an answer independently of the subject, from a single relevant aspect up to application in new contexts ([Biggs and Collis, 1982](#)). LRP uses it in four steps:

1. a single relevant aspect, correct;
2. several correct aspects, unconnected to each other;
3. the correct aspects, linked into a coherent explanation;
4. the coherent explanation, applied or extended to a new situation.

Below the first step is an absent answer or one unrelated to the question. The AI describes each step concretely for the generated question. **The steps are NOT grades:** the teacher maps them onto the local grading system

Two independent axes. The MCS level determines *what* is asked: the stage verified. SOLO determines *how* the answer is judged: its quality. This separation allows the same scheme to work at any level and in any subject.

Teacher's notes. An optional free-text field allows individualization: the desired form of the question, parts of the homework to avoid, particularities of the student or the class. The notes take precedence over the prompt's general rules, with one exception: each question has a single requirement. The exception comes from the NYU experience, where the voice agent frequently asked several questions at once and overwhelmed students ([Ipeirotis and Rizakos, 2026](#)).

6. Incentive Analysis

LRP does not try to make **cheating** impossible, but to make it **more costly than learning**. A protocol for real classrooms cannot control every student's behavior; it can only steer the incentives.

Random questioning. Every student knows they may be called on, but not when. The teacher's cost stays low (a few students per class), while the deterrent effect applies to the whole class.

Cheating strategies and their effect.

Student strategy	What happens	Assessment
Works at level 00 and submits the text without reading it	cannot answer the level-conditioned question	caught if selected
Works at level 00, then studies the text thoroughly	can answer	acceptable: they learned
Asks the AI for the teacher's likely questions and practices them	answers partially or well	acceptable: practicing answers is learning
Respects the required level	answers easily	the intended behavior

For the school, three of the four strategies lead to learning. The protocol does not distinguish between them, nor does it aim to.

The second question. If the first question misses what the student worked on, the student may ask for a second one, on the same stage. So that this option does not become a free exit, the protocol proposes two variants, at the teacher's choice:

- **grade cap:** a student who asks for a second question can obtain at most one step below the maximum of the local grading system;
- **increased probability:** a student who asks for a second question will (almost certainly) be selected for the next assignment as well. This variant does not depend on the grading scale, but may weigh more heavily on anxious students.

Goodhart's law. *When a measure becomes a target, it ceases to be a good measure.* The risk arises if the teacher or the school turns the results of LRP checks into a class performance indicator. LRP is proposed as a spot-check instrument, not as a reporting statistic.

7. Related Work and Positioning

LRP combines three existing directions: scales for AI use in assessment, the return of oral exams, and standard grading instruments. Its own contribution is the link between them: the oral question is derived from the declared level of division of labor with AI.

Scales for AI use. **AI Assessment Scale** (Perkins, Furze, Roe and MacVaugh, 2024) defines levels of permitted AI use in an assessment, from a complete ban to free use. **AI Friction Scale** (Beckwith, 2025) adds to each level a description of cognitive friction and signs of preparation (celt.uky.edu/.../ai-friction-scale.pdf). Both are rubrics for teachers and students. MCS turns a similar principle into a configurable behavior of the AI assistant, and LRP links it to verification.

Friction as intervention. Collins et al. showed experimentally that small frictions in a language model's interface reduce overreliance on it ([Collins et al., 2024](#)).

AI-conducted oral exams. Ipeirotis and Rizakos conducted 36 oral exams through a voice agent, at a total cost of \$15, i.e. \$0.42 per student. 70% of students felt the format tested their real understanding, while 83% found it more stressful than written exams ([Ipeirotis and Rizakos, 2026](#)). The system is technically advanced: the AI conducts and grades the exam. **LRP is simpler and differs in two respects: the teacher asks the question and gives the grade, and the question is conditioned on the declared level of work with AI,** not (only) on the content of the work.

Oral assessment as a response to AI. Several universities have introduced oral components to verify assignments. An ASCILITE study reports that some students who could not justify their reasoning orally showed signs of generative AI use (publications.ascilite.org/.../8608).

Criticisms. The literature contains serious objections to each component of LRP:

1. O'Sullivan argues that oral assessment, although hard to cheat with AI, can deepen existing pedagogical inequalities ([O'Sullivan, 2025](#)).
2. The validity of oral assessments is affected by examiner bias related to age and gender (publications.ascilite.org/.../8608), and student anxiety can distort the result (staff.universiteitleiden.nl/.../toolkit-assessment-methods--ai.pdf).
3. Furze opposes grading with generative AI because of inconsistency and bias ([Furze, 2024](#)).
4. Kohn describes rubrics as standardization tools that create an illusion of objectivity ([Kohn, 2006](#)), and Wilson shows how they can replace professional judgment ([Wilson, 2006](#)).

LRP addresses these criticisms partially **through design:** the AI does not grade, the rubric is optional (type A exists for teachers who reject rubrics), and the reference points are explicitly marked as support, not verdict. The

criticisms nonetheless remain valid relative to an ideal assessment; LRP is offered only as the **realistic** alternative to what the average teacher actually has: no verification at all, or AI detectors.

8. Limitations and Risks

The limitations below are known from the outset and must be checked in any application.

- **It does not verify slider use.** The protocol verifies understanding, not the process. A student may pass without having respected the required level, and one who respected it may lose out in the check if they have forgotten.

- **Discrimination between levels is assumed, not demonstrated.** It is not yet established whether a language model produces genuinely different questions for adjacent levels, for example 06 and 07.

- **Question difficulty varies.** Two students with the same homework may receive questions of different difficulty. Indicative content and levels of response reduce the variation without eliminating it.

- **Oral assessment has costs.** Anxiety, examiner bias, and class time consumed.

- **Choosing the level requires pedagogical judgment.** Without a guide, different teachers will choose different levels for similar assignments.

- **Unequal access.** The extension works only in compatible desktop browsers. The instructions-based variant in the AI assistant's app partially covers phones and tablets.

- **Age and platform terms.** AI assistants have minimum age limits, which differ from one platform to another and must be checked locally. For students below these limits, the only compatible option is parent-mediated use, on the parent's account.

- **Long instructions are followed imperfectly.** The teacher's prompt has several sections, and models may ignore rules located inside a long text. Structuring it into sections mitigates the risk but does not eliminate it.

9. Measurements Needed for an Empirical Evaluation of LRP

A **pilot application** of LRP should answer three questions, in this order:

1. **Does it discriminate between levels?** For the same homework excerpt, do questions generated at different levels differ in content, not just in wording? This can be tested without students: independent raters receive shuffled questions and guess the level.

2. **Does it discriminate between students?** Do students who respected the required level answer better than those who did not? A simple design: the same classes alternate assignments with and without LRP, each student serving as their own control.

3. **Does it change behavior?** Does use of the required level increase after random oral checks are introduced? A negative result on the first question would invalidate the protocol's central contribution and should be published just like a positive one. Results from applications, anonymized or not, can be submitted through the protocol page to be published for the benefit of other teachers.

10. Conclusions

LRP proposes a simple tool for the average teacher: students use AI for homework at a set level, and the teacher verifies understanding through random oral checks with questions conditioned on that level.

The protocol does not promise total control, nor does it seek it. It offers a cheap and fast way to maintain the link between the submitted homework and what the student knows, in a context where that link can no longer be assumed. Its real value can be established only through application in classrooms.

Appendix A. The Teacher's Prompt

The teacher fills in the input data and sends the entire text to the AI assistant.

INPUT DATA (*fill in before sending*)

Subject:

Age or year of study:

Assignment prompt:

Required MCS level (01-10):

Task type (A / B / C):

Teacher's notes (optional):

Excerpt from the student's homework:

ROLE

You generate oral verification questions for a teacher.

You do not assess and you do not grade. Any reference point you provide is support for the teacher's judgment, not a verdict.

CONTEXT: THE MCS MECHANISM

The student did the homework with an AI assistant configured at an MCS level (Mechanism of Cognitive Stimulation), between 01 and 10. The level determines which stages of the work belonged to the student and which stages the AI performed:

01: The AI gave the answer and explained the principle.

02: The AI gave the answer; the student made the decisions flagged by the AI.

03: The student set the goal or a constraint; the AI solved it.

04: The AI proposed 2-3 approaches; the student chose; the AI carried it out.

05: The student chose the approach and justified it; the AI carried it out step by step; the result was checked together.

06: The AI gave the framework and the first step; the student proposed the next steps.

07: The student attempted first; the AI gave hints and feedback.

08: The student led each key step; the AI guided and checked.

09: The student proposed the method and the solution; the AI made targeted corrections.

10: The student built the method, the solution and the verification; the AI only asked questions and flagged errors.

You do not know whether the student actually respected the level. You only know the required level. The question verifies whether the student now has the competence that

working at that level builds.

STAGE TO VERIFY, BY LEVEL

01: explaining the principle behind the solution.

02: one decision made and its effect on the result.

03: the goal or constraint and its influence on the result.

04: the existing alternatives and the reason for the choice.

05: justification of the approach and a spot check of the result.

06: continuing the solution from a given point.

07: redoing a step from the excerpt under modified conditions.

08: the sequence of key steps and how they connect.

09: applying the same method to a variant of the problem.
10: verifying the correctness of the result and the points of risk.

CONSTRAINTS (for any task type)

- Each question has a single requirement. No sub-questions.
- Anchored in the concrete content of the excerpt.
- Cannot be answered by reproducing the text from memory.
- Can be answered orally in at most 2 minutes, without access to the homework.
- Adapted to the age and subject given in the input data.
- The form of the question (verbal, computational or mixed) follows from the subject and the stage verified.
- Follow the teacher's notes, if any; they take precedence over the rules above, except for the rule of a single requirement per question.
- Language: the same as the homework.

TASK, BY TYPE

A (quick)

QUESTION: [one sentence]

B (standard)

QUESTION: [one sentence]

STAGE VERIFIED: [according to the level]

INDICATIVE CONTENT: [2-3 ideas that a correct answer might contain. The list is not exhaustive; any other correct answer is accepted.]

C (complete)

QUESTION: [one sentence]

STAGE VERIFIED: [according to the level]

INDICATIVE CONTENT: [as in B]

LEVELS OF RESPONSE: [describe concretely, for this question, what an answer looks like at each level]

- Level 1: a single relevant aspect, correct.
 - Level 2: several correct aspects, unconnected to each other.
 - Level 3: the correct aspects, linked into a coherent explanation.
 - Level 4: the coherent explanation, applied or extended to a new situation.
- Below level 1: absent answer, or unrelated to the question.

SECOND QUESTION: [same stage, different point in the excerpt]

The levels of response are not grades. The teacher maps them onto their own grading system.

For types A and B, a second question is generated only on request, on the same stage, from a different point in the excerpt, in the same format.

Appendix B. The Prompt Generator

To avoid editing the prompt by hand, the protocol page (cognitiveslider.org/lrp) includes a form that assembles the complete prompt into a single box, with a copy button.

1. **The MCS level and the task type** are chosen from **drop-down lists**, with a short description of each option.
2. The prompt is assembled **only in the teacher's browser**. **The data entered is not sent** to any server.
3. The teacher can pin the subject, year of study, MCS level and task type individually. Pinned values are kept **only in their browser** (*localStorage*) and **can be deleted at any time**.
4. The assignment prompt, the student's excerpt and the notes **cannot be pinned**, because they may contain data about students.

References

- Bastani, H. et al. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *PNAS*, 122(26). [doi:10.1073/pnas.2422633122](https://doi.org/10.1073/pnas.2422633122)
- Beckwith, J. (2025). *AI Friction Scale*. [PDF](#)
- Biggs, J. and Collis, K. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy*. Academic Press. [description](#)
- Collins, K. M. et al. (2024). Modulating Language Model Experiences through Frictions. [arXiv:2407.12804](https://arxiv.org/abs/2407.12804)
- Fernandes, D. et al. (2026). AI makes you smarter but none the wiser. *Computers in Human Behavior*, 175. [doi:10.1016/j.chb.2025.108779](https://doi.org/10.1016/j.chb.2025.108779)
- Furze, L. (2024). Don't use GenAI to grade student work. [article](#)
- Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, 15(1). [doi:10.3390/soc15010006](https://doi.org/10.3390/soc15010006)
- Ipeirotis, P. and Rizakos, K. (2026). Scalable and Personalized Oral Assessments Using Voice AI. [arXiv:2603.18221](https://arxiv.org/abs/2603.18221)
- Kohn, A. (2006). The Trouble with Rubrics. *English Journal*, 95(4). [article](#)
- Lee, H.-P. et al. (2025). The Impact of Generative AI on Critical Thinking. *CHI '25*. [doi:10.1145/3706598.3713778](https://doi.org/10.1145/3706598.3713778)
- O'Sullivan, J. (2025). Oral assessment won't save higher education from generative AI. [article](#)
- Perkins, M., Furze, L., Roe, J. and MacVaugh, J. (2024). *AI Assessment Scale*. aiassessmentscale.com
- Stan, A. (2026). *Open Letter to Those with Their Hand on the Slider of the Future*. cognitiveslider.org
- Leiden University. *Assessment methods & AI toolkit*. [PDF](#)
- Wilson, M. (2006). *Rethinking Rubrics in Writing Assessment*. Heinemann. [book page](#)
- *ASCILITE* study on oral assessment and generative AI. [PDF](#)