Creative Education, 2026, 17(8), 1556-1586
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

AI-DRS: A Dialogic Generative-AI Instructional Model for Scientific Reasoning in Primary Education

Michail Kalogiannakis[1*], Theodoros Spasopoulos[1], Nikos Papakonstantinou[1], Apostolos Xenakis[2]

[1]Department of Special Education, University of Thessaly, Volos, Greece
[2]Department of Digital Systems, University of Thessaly, Larissa, Greece
Email: *mkalogian@uth.gr

How to cite this paper: Kalogiannakis, M., Spasopoulos, T., Papakonstantinou, N., & Xenakis, A. (2026). AI-DRS: A Dialogic Generative-AI Instructional Model for Scientific Reasoning in Primary Education. *Creative Education, 17,* 1556-1586.
https://doi.org/10.4236/ce.2026.178091

Received: June 24, 2026
Accepted: August 22, 2026
Published: August 25, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Cultivating scientific reasoning in primary education involves not just making measurements, but also turning them into evidence and explanations. Physical-computing activities often do not support this process, and using generative artificial intelligence (GenAI) as a tool for answers can skip this step. This article suggests a teaching method for primary education that limits GenAI's role so it can support pupils' reasoning and then step back. In our research-and-development design, we created the AI-DRS (AI-supported Dialogic Reasoning Scaffolding) Physical Inquiry Model. This model pairs a Nezha-micro:bit soil-moisture investigation with sensor data generated by pupils. It follows a Claim-Evidence-Reasoning framework and includes a dialogic control layer based on three principles: grounding in data, ensuring proper epistemic oversight, and being responsive to pupils. The model incorporates fading based on established criteria, teacher-led interventions, and specific protections for each child. The article specifies the scientific target of the investigation, the sensor-calibration and data-quality rules, the assignment and balancing of inquiry variables, the technical configuration and rule precedence of the system, a teacher escalation protocol with explicit stop rules, and a cluster-aware analysis plan. Next steps include expert validation and a controlled pilot study. The model has a seven-phase process, an architecture that responds to pupil input, a scoring system, and a coding scheme that is sensitive to how pupils engage. It separates the tool's dialogic moves from pupil responses, allowing for individual analysis. This proposal is theoretical, with no empirical results presented. It establishes a structured approach where GenAI can ask follow-up questions while the teacher maintains control over the knowledge. Validating the model's effectiveness and understanding how it works is the essential next step.

Keywords

Generative Artificial Intelligence, Micro:bit, Claim-Evidence-Reasoning, Scientific Reasoning, Primary Education

1. Introduction

1.1. Background of the Study

Scientific literacy in primary education rests on a skill that is simple to describe but hard to teach: reasoning from evidence. Children must distinguish observation from interpretation, make testable claims, and select data that support them. The most demanding step is explaining why that data supports the claim through a scientific mechanism. In primary classrooms, pupils often reach the conclusion but struggle with the explicit mechanistic link between evidence and claim (Diola et al., 2025; McNeill et al., 2006).

Physical computing eases this challenge. Sensors let children build systems that interact with the physical world and return continuous real-time measurements, turning abstract phenomena into observable data (Ntourou et al., 2021; Przybylla & Romeike, 2014; Cao et al., 2021). Reliable data, however, do not produce interpretation on their own: the critical step is moving from recording a value to reading a pattern as evidence of a mechanism, and without support, pupils find this hard (Lin et al., 2023). GenAI promises personalized support, timely feedback, and new forms of assessment (Yan et al., 2024), but used as an answer provider it performs the reasoning the learner should do, encouraging over-trust and weakening independent thinking (Cooper, 2023; Lee et al., 2025). The question for primary science is therefore not whether GenAI will be present, but which instructional design can make it support reasoning rather than replace it.

1.2. Problem of the Study

Research on the micro:bit has focused mainly on technical features, motivational benefits, and approaches to teaching programming, and less on whether such activities support scientific thinking or conceptual understanding (Sentance et al., 2017; Kalelioğlu & Sentance, 2020). Completing a physical-computing task does not demonstrate understanding of the phenomenon, the variables, or the link between measurement and scientific model: a device that works is not a pupil who thinks. Similarly, many educational chatbots ask generic questions and give feedback without requiring pupils to draw conclusions from their own data, and without criteria for when to advance, persist, hint, or withdraw.

Two gaps emerge. Physical computing generates rich data but does not ensure scientific reasoning; GenAI generates rich dialogue but can easily detach from the learner's data and substitute for the learner's own thinking. This study treats the two gaps as complementary and addressable through a single intervention, provided the dialogue stays tied to pupils' own measurements and follows explicit

epistemic and developmental rules.

1.3. State of the Art

Dialogic teaching views learning as a process in which ideas are shared, examined, and progressively refined through mutual interaction (Alexander, 2018; Muhonen et al., 2016; Vrikki et al., 2019). Unlike static prompts, generative artificial intelligence (GenAI) dialogue systems can adapt to how learners are thinking, which makes them a promising support tool when they advance inquiry through questions rather than immediate answers (Kalogiannakis et al., 2025; Sotiropoulos et al., 2026; Spasopoulos et al., 2025; Tang & Putra, 2026). LLM-based platforms have also been developed to help educators design computational-thinking, AI, and STEM activities, illustrating the wider potential of generative systems as pedagogical infrastructures (Xenakis et al., 2025).

Customized chatbots can support reasoning and argumentation in secondary science (Tang & Putra, 2026), and recent theoretical work casts GenAI as a conversational partner rather than a source of information (Wegerif & Casebourne, 2026). Neither specifies an implementation for primary-aged children: a developmentally appropriate method grounded in children's own sensor data, with rules for when the system should question, prompt, or support, and when that support should be withdrawn. Scaffolding theory supplies the missing element. A scaffold is the support that enables a learner to solve a problem beyond their current ability (Wood et al., 1976), and its defining features are contingency, fading, and the transfer of responsibility to the learner (van de Pol et al., 2010); permanent mediation is not effective support, even when the immediate answer is correct. Work on scientific explanation stresses reducing written aids as skill develops (Masters & Docktor, 2022; McNeill et al., 2006), and studies of AI-supported scaffolding emphasize continuous diagnosis and progressive withdrawal rather than answer provision (Bai et al., 2026; Cai et al., 2025; Wan Hamedi et al., 2025).

Three further findings clarify the developmental stakes. Young children may place considerable and largely uncritical trust in conversational agents (Movahed & Martin, 2025). AI literacy must be deliberately cultivated rather than assumed to emerge from technology use (Ng et al., 2024). And chatbot effects are, meta-analytically, smaller for younger learners, with novelty-related gains that may fade (Wu & Yu, 2024). Syntheses of GenAI-supported learning agree: the technology's educational potential cannot be separated from its effects on cognition, metacognition, and learner agency (Yan et al., 2024).

Consequently, general-purpose chatbot designs cannot simply be transferred to primary education; they require developmentally appropriate constraints, scaffolding, and safeguards.

1.4. Research Gap and Objective

The closest existing demonstration, a customized chatbot supporting reasoning and argumentation in secondary science (Tang & Putra, 2026), leaves three issues

unaddressed for primary education: dialogue grounded in pupils' own sensor data rather than curricular text, explicit criteria for fading support, and safeguards for children who may over-trust conversational agents. Effectiveness is defined here not by whether a pupil completes a conversation but by whether the pupil can subsequently produce an independent Claim-Evidence-Reasoning (CER) explanation (McNeill & Krajcik, 2012) without AI. The aim is to develop the AI-DRS (AI-supported Dialogic Reasoning Scaffolding) Physical Inquiry Model and prepare it for validation, treating the design itself as the research contribution.

The study is guided by a single design question: which design principles and architecture can connect GenAI dialogue to primary pupils' own sensor data so that it supports, and then deliberately fades, independent evidence-based reasoning? The evaluative questions that follow concern whether the model improves independent CER performance relative to an equivalent static scaffold, how it affects pupils' use of their own data and the trajectory of scaffolding, whether pupils articulate synthesis themselves, whether reasoning transfers without AI, and how perceived usefulness and calibrated trust relate to reasoning quality. These form the protocol of the planned pilot and are not addressed empirically here.

2. Method

2.1. Type and Design

The study uses a Research and Development (R&D) design with two sequential parts: (a) developing the intervention model and (b) planning its expert validation and subsequent pilot evaluation. This article reports only the first part; the second is set out as a protocol for future work. The study is deliberately not presented as an efficacy trial: the planned pilot examines feasibility, implementation fidelity, preliminary indications of effect, and above all, the links among scaffolding, data use, and reasoning.

Development is organized into seven R&D stages:

1) needs analysis through curriculum documents, research literature, and the authors' teaching experience, with no participant data;

2) mapping of curriculum, target concepts, and likely misconceptions;

3) development of the inquiry task, the AI-DRS prompt, the worksheet, and the teacher guide;

4) expert validation;

5) a small-group technical and usability test;

6) revision of the package; and

7) pilot implementation with process evaluation.

The first three stages are complete, and their output, the model in Section 3, constitutes this article's contribution; the remainder are reported as protocol, and Sections 2.2 - 2.4 set it out in full for replicability, describing planned rather than collected data. Because the intervention involves GenAI use with children, it assumes a school-managed environment, data minimization, parental consent and pupil assent, pseudonymized chat logs, defined retention rules, and continuous

teacher oversight. The system states that it can be wrong, asks pupils to check its suggestions against their own observations, and routes uncertain or problematic cases to the teacher. The planned pilot will be submitted to the Research Ethics Committee of the University of Thessaly before any data collection (Petousi & Sifaki, 2020).

The design must also address how children may lawfully interact with GenAI. Mainstream consumer chatbots are not permitted for this age group: OpenAI's terms require users to be at least 13, with parental consent for minors (OpenAI, 2025), and Anthropic's require users to be at least 18 (Anthropic, 2025). The model, therefore, assumes mediated access. Pupils hold no accounts and use no consumer interface; the dialogue runs through a school-controlled application operating under the teacher's institutional account and the selected provider's application programming interface (API). That application logs pseudonymized interaction data inside the school environment and transmits only data-minimized, non-identifying inputs through the API, under approved data-processing, security, and retention arrangements.

2.2. Data and Data Sources

A sample of about 40 to 60 Grade 5 or Grade 6 pupils is proposed, with 20 to 30 per condition. Because pupils build and measure in groups of three to four and groups are the unit of assignment, this corresponds to approximately five to ten groups per condition, a figure that governs the analysis plan in Section 2.4. Pupils work in groups during construction, but the pre-test, post-test, and transfer measures are administered individually. The planned data sources are the individual pre- and post-CER tasks, an individual transfer task completed without AI, pseudonymized chat logs, teacher observation logs, technical and fidelity records, a short perceived-usefulness and calibrated-trust questionnaire, and the expert-validation ratings collected before the pilot.

The intervention is built on the Nezha Inventor's Kit for micro:bit (**Figure 1**). Because the assessed comparison requires both soil conditions to be set up and measured simultaneously within each group (Section 2.2.3), every group is equipped with two complete measurement channels, one per container: two BBC micro:bit units, two soil-moisture probes, and LED indicators. This exceeds the standard kit, which contains a single soil-moisture sensor, so a second kit, or a second micro:bit and probe, is required per group; this is stated explicitly as a resource condition of the design rather than assumed.

The materials support a Smart Plant Investigation in which pupils build a soil-moisture monitoring system and examine how soil type affects the change in soil moisture over time. Every group investigates the same focal comparison under the same controlled conditions, so that datasets and reasoning demands are equivalent across groups and conditions. After calibration and a stability check on each channel, pupils record the container, elapsed time, calibrated readings, quality flags, and their observations.



Figure 1. Components of the Nezha Inventor's Kit used in the proposed Smart Plant Investigation. Source: Authors' photograph.

The soil-moisture probe is the kit's own sensor, which keeps the build within the kit's wiring conventions and MakeCode blocks; components such as a BME280 sensor or an OLED display are not part of the standard configuration (ELECFREAKS, n.d.). Duplicating the micro:bit and the probe, as described above, is the only departure from the standard configuration.

2.2.1. Inquiry Question, Target Mechanism, and Permissible Causal Claim

Claim-Evidence-Reasoning (CER) responses can be scored consistently only if the investigation has one question, one expected mechanism, and one clearly bounded causal claim that applies to every group. The Smart Plant Investigation is therefore built around a single focal comparison rather than group-selected variables. The operational inquiry question is: how does soil type (sandy soil versus potting compost) affect the rate at which soil-moisture readings decrease over 40 minutes, when the initial water volume, container, soil mass, probe placement, bench position, light exposure, and room temperature are held constant?

The expected empirical pattern is a steeper decline of the calibrated Moisture Index in sandy soil than in compost. The target mechanism, expressed at a level appropriate for Grade 5-6, is that sandy soil consists of larger particles with larger pore spaces and little organic matter, so water drains through it faster and is exposed to the air across a larger internal surface, whereas the finer particles and organic matter in compost hold water against drainage and evaporation. A full-credit reasoning statement links a stated numerical difference in the rate of decline to this particle-size and pore-space account; a partial-credit statement identifies the difference but restates the data instead of explaining it, as in "it dried faster

because the numbers went down faster".

The permissible causal claim is deliberately bounded. Because soil type was manipulated while the remaining conditions were held constant, pupils may claim that, in this investigation and within the 40-minute window, the type of soil caused the difference in the rate of moisture loss. They may not claim which soil is better for growing plants, extend the claim to plant health or growth, read the index as an absolute quantity of water, or make causal claims about light, temperature, or container size. Table 1 states this target explicitly so that raters, the teacher, and the AI-DRS system apply the same scientific standard: it is supplied to the system as a fixed reference in every dialogue turn (Section 3.4) and it governs the Mechanism dimension of the rubric (Section 3.7).

Table 1. Scientific target of the smart plant investigation.

Element	Specification
Operational inquiry question	How does soil type (sandy soil vs. potting compost) affect the rate at which soil-moisture readings decrease over 40 minutes, when all other conditions are held constant?
Manipulated variable	Soil type, at two levels: sandy soil and potting compost. Both levels are set up and measured simultaneously within every group, on two separate probes.
Outcome measure	Calibrated Moisture Index (MI, nominally 0 - 100; the index is unbounded and is logged unclipped, and only the plotted value is clipped to −5 to 105) recorded every 5 minutes for 40 minutes on each of the two containers; derived measures are the total decline (MI at t0 minus MI at t40) and the mean rate of decline (MI points per 10 minutes).
Controlled variables	Container type and volume, dry soil mass (±5 g), initial water volume (100 ml), probe depth and position, bench position and light exposure, room temperature, and time of day.
Expected pattern	A steeper decline in sandy soil than in compost across the measurement window.
Target mechanism (pupil level)	Sandy soil has larger particles and larger pore spaces and little organic matter, so water drains faster, and more of it is exposed to the air; the finer particles and organic matter in compost hold water against drainage and evaporation.
Permissible causal claim	"In this test, the type of soil caused the difference in how fast the soil lost water over 40 minutes", supported by at least two specific readings from each container.
Claims outside scope	Which soil is better for plants; effects on growth or plant health; absolute water content; behavior beyond 40 minutes or after re-watering; causal claims about light, temperature, or container size.
Anticipated misconceptions	Reading the sensor value as an absolute amount of water; attributing the decline to the plant "drinking" it; treating a single reading as a trend; equating a faster decline with a "worse" soil.

2.2.2. Sensor Calibration, Measurement Schedule, and Data-Quality Rules

The soil-moisture sensor returns an uncalibrated analog value, so the same physical condition can produce different readings across sensors, sessions, and probe placements. Interpretation is therefore preceded by a fixed two-point calibration and a stability check, and the dataset is screened with rules defined in advance rather than negotiated during the dialogue.

Calibration is carried out for each of a group's two probes at the start of every session and repeated whenever a probe is removed and re-inserted: a dry reference in air and a wet reference at the marked depth in tap water, each the mean of five

readings after a settling period (Table 2). Readings are converted in MakeCode to a Moisture Index, MI = 100 × (R − R_dry)/(R_wet − R_dry). The index is unbounded: the unclipped value is logged and screened for quality, and only the plotted value is clipped to −5 to 105, so that an out-of-range reading remains detectable by the range rule. Each channel yields its own pair of calibration constants; both pairs are recorded on the group worksheet and stored with the dataset, labeled by container.

The acceptance criteria and the action taken when they are not met are set out in Table 2: a reference is accepted only when the five readings span no more than 2% of full scale, and a probe that fails three attempts, or whose calibration span falls below 150 raw units, is replaced before the investigation proceeds. After the standardised watering at t = 0, MI is recorded on both channels every 5 minutes for 40 minutes, giving nine time points per container; each value is the median of three consecutive readings taken 5 seconds apart, which suppresses single-sample noise without requiring statistical treatment by pupils. Probes remain at a fixed depth of 5 cm and at least 2 cm from the container wall, marked with tape; any displacement is recorded as an event.

The investigation is timetabled across three 45-minute lessons, set out in the final row of Table 2. Lesson 2 is the tightest, since a 40-minute window leaves no margin in a 45-minute slot; it is therefore timetabled as a double period where the school allows one, and otherwise calibration is completed at the end of lesson 1 so that lesson 2 begins with watering at t = 0. If neither is workable, the schedule is shortened to 30 minutes and seven time points, but only as a study-level decision taken for the whole pilot before data collection, never per group, so that all groups produce datasets of identical structure; the adopted schedule and its dependent thresholds are then reported. The 40-minute, nine-point schedule is the reference version used throughout this article. Lesson 2 deliberately does not carry the dialogue: about four minutes between readings allows graphing and the claim and data-grounding moves but not the full contingent exchange, so conceptual bridging, fading, and transfer occur in lesson 3 from a closed dataset.

Table 2. Calibration and measurement procedure.

Step	Procedure	Acceptance criterion	Action if not met
Dry reference (R_dry)	Each probe clean, dry, in air; 60 s settling; mean of five readings at 10 s intervals.	Spread of the five readings ≤ 20 raw units (2% of full scale).	Wait 60 s and repeat, up to three attempts; then replace the sensor.
Wet reference (R_wet)	Each probe inserted to the marked depth in tap water, electronics above the waterline; 60 s settling; mean of five readings.	Spread ≤ 20 raw units.	As above; a probe failing three attempts is withdrawn.
Span check	Compute |R_wet − R_dry| and record it on the worksheet.	Span ≥ 150 raw units.	Re-seat the probe and check the connector; if unchanged, replace the unit.
Conversion	MI = 100 × (R − R_dry) / (R_wet − R_dry), computed on each micro:bit. The unbounded index is logged; only the plotted value is clipped to −5 to 105.	Both pairs of calibration constants logged with the dataset and labeled by container.	The dataset is not interpreted until the constants are recorded.

Continued

Measurement schedule	Reading on both channels every 5 minutes for 40 minutes after watering at t = 0; each value is the median of three readings taken 5 s apart.	Nine time points per container under the reference schedule, with no missing points.	A missing point is re-taken at the next interval and marked as delayed.
Probe placement	Fixed depth of 5 cm, at least 2 cm from the container wall, position marked with tape.	Probe not moved between t = 0 and t = 40.	Displacement is logged as an event; subsequent readings are flagged.
Session scheduling	Three 45-minute lessons: (1) orientation, prediction, fair test, construction, programming, calibration; (2) re-calibration, watering at t = 0, the 40-minute window; (3) CER dialogue, fading, transfer task.	Lesson 2 starts from completed calibration and ends with nine time points per container.	A double period where available; a shortened 30-minute, seven-point schedule may be adopted for the whole pilot before data collection, never for single groups. An unfinished run is not interpreted; lesson 3 uses the class reference dataset.

Five rules identify unreliable readings before any interpretation begins (Table 3). A single flagged point is re-taken once and, if the flag recurs, is excluded from the evidence base while remaining in the log. A run that fails the drift rule, or a container with more than two flagged points out of nine, is not interpreted at all: the system halts interpretation and refers the group to the teacher, which is the operational trigger for rule R8 in Table 8. The group then completes the CER task using a validated class reference dataset, so that a hardware fault does not cost the pupils the reasoning task. These definitions give the system's unreliable-data category a concrete referent instead of leaving it to real-time judgment by a generative model.

Table 3. Rules for identifying unreliable readings.

Rule	Trigger	Interpretation	Consequence
Range	The unbounded index falls outside −5 to 105 MI, screened before the plotted value is clipped.	Calibration or contact failure.	Point discarded and re-taken once; a recurrence flags the container.
Jump	A change greater than 15 MI points between consecutive 5-minute points after t = 5, with no logged event.	Probe movement or intermittent contact. The first interval (t = 0 to t = 5) is exempt because rapid drainage after watering is the phenomenon under study; the threshold is set per soil condition from the class reference dataset.	Point re-taken once; if it recurs, the point is excluded from the evidence base.
Flatline	An identical raw value for four or more consecutive readings.	Disconnection or a frozen input.	Hardware check by the teacher before measurement resumes.
Drift	A post-session dry check differing from R_dry by more than 25 raw units, or by more than 5% of the calibration span, whichever is larger.	Sensor drift across the session, judged against a floor that exceeds the calibration noise accepted above.	The whole run is flagged unreliable and is not interpreted.
Run level	More than two flagged points in a container, out of the nine recorded under the reference schedule.	The dataset cannot support an evidence claim.	Interpretation halted, teacher check, and the class reference dataset is used for the CER task.

The class reference dataset is not improvised during the lesson. It is collected

by the research team in a pre-pilot, with at least six complete runs per soil condition, using the same kit, soils, containers, calibration procedure, and 40-minute schedule as the pupils, and screened against the same five rules. Its pooled runs fix the per-condition jump thresholds used in Table 3, provide the substitute dataset for a run that cannot be interpreted, and establish that the two soils are distinguishable within the measurement window. Because the substitute data follow the same protocol, a group working from them faces the same CER task; every substitution is recorded as a fidelity deviation and reported with the results.

2.2.3. Assignment and Balancing of Inquiry Variables

Allowing each group to select its own variable would produce datasets of different shapes and reasoning tasks of different difficulty, confounding any comparison of CER performance between conditions. Here the focal variable is identical for all groups, and both of its levels are investigated within each group: every group sets up one sandy-soil container and one compost container and measures them simultaneously on its two channels. Every group therefore produces a dataset of the same structure, two conditions by nine time points, and faces the same reasoning demand. Light exposure, initial water amount, and container surface area are retained only as optional extensions after the assessed CER task, recorded as process variables and never used in a between-condition comparison.

Assignment proceeds in three steps. First, pupils are ranked within class on the pre-test CER total and allocated to mixed-attainment groups of three to four, so that group mean prior attainment is comparable. Second, groups rather than pupils are randomly assigned to the AI-DRS or the static-scaffold condition, using block randomization with class as the blocking factor and a computer-generated sequence prepared by a researcher not involved in delivery; allocation is concealed until the pre-test has been scored.

The group is thus both the unit of assignment and of delivery, which is the basis of the analysis in Section 2.4. Third, equipment is counterbalanced: probes, micro:bit units, and bench positions are rotated across conditions and across the two soils within each group, so that no device or location is systematically associated with one condition or one soil.

Standardization of the materials is specified in Table 4. Balance is verified before analysis by reporting, separately by condition, the distribution of pre-test scores, soil batches, sensor units, bench positions, ambient temperature and relative humidity, and prior micro:bit and GenAI experience. Any imbalance is reported descriptively rather than adjusted away: with only five to ten groups per condition, randomization alone cannot be relied upon to produce equivalence, and the comparisons are interpreted accordingly.

Table 4. Standardization and assignment of the inquiry variables.

Variable	Status in the design	Specification	Balancing procedure
Soil type	Manipulated, within group	Sandy soil and potting compost, equal dry mass (±5 g).	Both levels in every group; one batch per soil type for the whole pilot.

Continued

Initial water volume	Controlled	100 ml delivered to each container with the same syringe at t = 0.	Identical instruction and equipment; volume recorded on the worksheet.
Container	Controlled	Identical pots: same volume, surface area, and drainage.	Supplied pre-assembled and checked before the session.
Probe depth and position	Controlled	5 cm depth, at least 2 cm from the wall, marked with tape.	Checked by the teacher before t = 0.
Bench position and light	Controlled in the assessed task	Same room and bench, no direct sunlight.	Positions rotated across conditions.
Temperature and humidity	Recorded covariate	Recorded once per session.	Both conditions run in the same room within the same two-hour window; cross-condition exposure is logged and reported (Section 2.2.4).
Probes and micro:bit units	Recorded and counterbalanced	Two probes and two micro:bit units per group, one per container; unit identifier logged with each container.	Units rotated across conditions and across the two soils within each group.
Group composition	Assigned	Mixed-attainment groups of three to four, stratified on pre-test CER.	Block randomization of groups to condition, with class as the block.
Light, water amount, surface area	Manipulated only in the optional extension	Held constant throughout the assessed CER task; varied only in the optional extension that follows.	Excluded from all between-condition comparisons.

2.2.4. Comparison Conditions

The two conditions are identical in every respect except the form of analysis support: the experimental groups receive the adaptive AI-DRS dialogue, the active control groups an equivalent static worksheet with the same questions in a fixed order. Hardware, phenomenon, dataset, time on task, and the individual final assessment are the same in both (Table 5). One source of contamination cannot be removed at this scale. Because both conditions run in the same room within the same two-hour window (Table 4), control groups can see and overhear the AI-DRS groups, and any diffusion of dialogic moves would attenuate the observed difference. Separating the conditions by room or session would confound condition with setting, so cross-condition exposure is instead logged and reported as a limitation.

Table 5. Comparison conditions.

Element	Experimental group	Active control
Hardware and phenomenon	Nezha-micro:bit Smart Plant inquiry	Identical
Dataset	Pupil-generated sensor data	Own group-generated dataset, identical protocol
CER questions	Adaptive AI-DRS dialogue	Equivalent static worksheet
Time on task	Equivalent	Equivalent
Teacher support	Recorded	Recorded
Final assessment	Individual, without AI	Individual, without AI

2.3. Data Collection Technique

Data will be collected across the seven-phase learning sequence detailed in Section 3. The system captures pupils' interactions as pseudonymized chat logs, their measurements as datasets and graphs, and their explanations as written CER responses, while the teacher completes a structured observation log throughout. Equivalence and fidelity indicators are recorded: prior science knowledge, baseline CER performance, prior micro:bit and GenAI experience, time on task, absences, technical problems, teacher interventions, and number of AI turns. A subset of lessons is reviewed against a fidelity checklist by a second observer.

2.4. Data Analysis

In the planned pilot, two independent raters will score at least 25% of the responses. Weighted Cohen's kappa will be computed for the ordinal rubric dimensions and an intraclass correlation coefficient for the total CER score, with disagreements resolved through documented consensus. Chat logs will be examined using a scheme that classifies the system's moves as Socratic questioning, data grounding, conceptual bridging, guided choice, explicit correction, or metacognitive prompting, and pupil responses as full, partial, none, or not reached. Synthesis is recognized only when the pupil articulates it; accepting an AI-generated summary does not count as independent reasoning.

The analysis plan follows from the unit of assignment. Because groups of three to four pupils are randomized and taught as a unit while outcomes are measured individually, pupils are nested within groups, and the planned sample corresponds to only five to ten groups per condition. An individual-level ANCOVA treating pupils as independent observations is therefore not an appropriate primary analysis, where the intraclass correlation is positive, standard errors are understated, and the associated p-values are not interpretable, and reporting an intraclass correlation alongside an uncorrected model does not remedy this (Hedges, 2007). Multilevel modeling is not a workable alternative at this scale either, because variance components estimated from fewer than roughly ten clusters per arm are unstable and their standard errors are biased downwards (McNeish & Stapleton, 2016).

The primary analysis is therefore specified at cluster level and is descriptive rather than confirmatory. For each group, the mean post-test and mean pre-test CER score are computed, and the between-condition contrast is the difference in group mean post-test scores adjusted for group mean pre-test scores. Every group mean is displayed individually, so that the small number of clusters is visible rather than hidden behind an aggregate. The effect size is a cluster-adjusted standardized mean difference using the total between-plus-within variance in its denominator, reported with the intraclass correlation on which it depends (Hedges, 2007). The primary inferential procedure is an exact permutation test on the groups' condition labels, because with five to ten clusters per arm it is the only exact procedure available. At the lower end of that range, five groups per arm, the test cannot return a two-sided p-value below about .008 and has negligible power

against any plausible effect; it is therefore reported as a descriptive statement of how extreme the observed contrast is among the possible label assignments, not as a test of efficacy. Interval estimates use a wild cluster bootstrap-t, developed precisely because pairs-resampling and percentile-type bootstraps behave poorly with few clusters; a bias-corrected and accelerated bootstrap resampling of whole groups is deliberately not used, for the same reason (Cameron et al., 2008). Both procedures are conservative at this number of clusters and the resulting intervals will be wide, which is accepted as the cost of correct inference rather than concealed behind a nominally narrower pupil-level interval.

Pupil-level ANCOVA is retained only as a secondary, illustrative analysis, with cluster-robust (CR2) standard errors and Satterthwaite degrees of freedom, designed for this situation of few clusters (Bell & McCaffrey, 2002); no pupil-level p-value is treated as evidence of effectiveness. The intraclass correlation is reported as an interval, not a point estimate, since with at most ten clusters per arm its confidence interval is necessarily wide, and it is accompanied by a sensitivity table giving the design effect, $\text{Deff} = 1 + (m - 1) \times \text{ICC}$, and the effective sample size, N/Deff, across a plausible range: with groups of three to four, an intraclass correlation of .05, .10, or .20 reduces an effective sample of 60 pupils to about 53, 48, or 40. This bounds the number of clusters a subsequent cluster-randomized trial would require, where the number of groups rather than of pupils determines power; it is a planning range, not an estimate a pilot of this size could deliver. The analysis plan, including the stopping rules for the process measures, will be pre-registered. Table 6 summarizes the unit of analysis, estimand, and inference for each outcome.

Table 6. Analysis plan: outcomes, units of analysis, estimands, and inference.

Outcome	Unit of analysis	Estimand	Inference	Status
Post-test CER	Group (5 - 10 per condition)	Difference in group mean post-test scores, adjusted for group mean pre-test scores.	Exact permutation test on group labels (primary); wild cluster bootstrap-t for interval estimates.	Primary, exploratory
Transfer task without AI	Group	Difference in group means on the individually administered transfer task.	As for the primary outcome.	Key outcome, exploratory
Post-test CER	Pupil	ANCOVA-adjusted difference with pre-test as covariate.	CR2 cluster-robust standard errors with Satterthwaite degrees of freedom; p-values not reported as evidence.	Secondary, illustrative
Clustering parameters	Group	Intraclass correlation as an interval; design effect and effective sample size across a sensitivity range (0.05, 0.10, 0.20).	Interval estimate with an explicit sensitivity table; no point estimate is relied upon.	Planning range for a future trial
Chat-log process codes	Turn and pupil	Frequencies of system moves and of pupil uptake.	Descriptive statistics and joint displays.	Descriptive
Rater agreement	Response	Weighted Cohen's kappa and the intraclass correlation for totals.	Point estimates with confidence intervals.	Quality control
Fidelity and feasibility	Session and group	Completion, time on task, technical failures, teacher interventions, flags raised.	Descriptive statistics.	Feasibility

The qualitative analysis will focus on patterns of CER error and will use thematic analysis for explanations, integrating turn-level chat-log coding with comparative cases of full, partial, and zero uptake in joint displays that connect CER gain, data-grounding frequency, bridge uptake, engagement, and transfer.

3. The AI-DRS Physical Inquiry Model

The development strand produces the AI-DRS Physical Inquiry Model, a proposed intervention in which GenAI acts as a dialogic scaffold that helps primary pupils move from observation to evidence-based scientific reasoning and then steps back. The sections below set out its reasoning progression, control principles, seven-phase sequence, response-contingency architecture, implementation specification, escalation protocol, fading logic, assessment instruments, and alignment with the Technological Pedagogical Content Knowledge (TPACK) framework. Figure 2 illustrates its overall logic.

3.1. From Measurement to Scientific Explanation

Sensors convert physical states into numerical values, but their scientific meaning comes from the relations among values, conditions, and theoretical ideas. Pupils must decide which data are relevant, whether a measurement is trustworthy, which patterns are present, and what mechanism might explain them. The model foregrounds four practices: generating data, checking data quality, selecting evidence, and explaining mechanistically. It organizes them through a CER progression (McNeill & Krajcik, 2012) adapted for the primary years: Observe/Describe, Predict, Claim, Evidence, Reasoning, and Reflect and Transfer.

At the heart of the model is a dialogic control layer that responds to the pupil's input. The system advances the pupil's thinking through targeted questions instead of presenting a ready-made explanation: the interaction starts with observation, asks for comparisons, requires a claim, returns the pupil to the actual measurements, and only then prompts for mechanistic reasoning. Its value depends on the quality of the pupil's responses, not the number of exchanges. Three principles govern the interaction. Data grounding requires pupils to cite specific values, changes, or patterns from their own data and rejects vague statements such as "the soil dried out more". Epistemic gatekeeping makes the move from claim to evidence conditional on a testable claim, the move from evidence to reasoning conditional on relevant data, and completion conditional on a mechanism stated in the pupil's own words.

Summaries produced by the system are never counted as the pupil's own synthesis. Contingent responsiveness determines the next move from the quality of the response, classified as adequate and independent, partially adequate, off-target, or scientifically inaccurate, for instructional routing only. Figure 2 shows these principles as a layer beneath the reasoning progression, with AI support decreasing as reasoning becomes more independent.

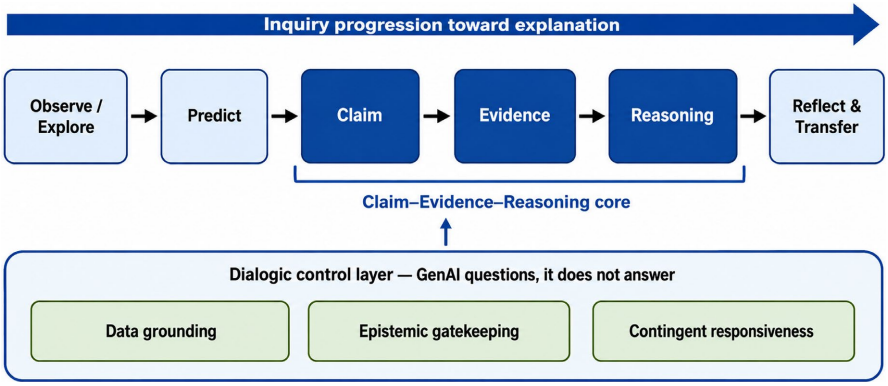


Figure 2. The AI-DRS reasoning progression and its dialogic control layer.

3.2. The Seven-Phase Learning Sequence

The intervention is enacted through seven phases, each pairing a pupil activity with a defined role for the AI-DRS system or the teacher and yielding a distinct piece of evidence for analysis (Table 7). The system does not answer during orientation, offers no correction during prediction, and enforces fair-test criteria before construction begins.

Table 7. The seven phases of the AI-DRS Physical Inquiry Model.

Phase	Pupil activity	AI-DRS/teacher role	Evidence produced
1. Orientation	Confront a real watering problem	Teacher activates prior ideas; no AI answer	Initial conceptions
2. Prediction	Predict the pattern and give an initial rationale	AI requests a clear prediction, not a correction	Prediction statement
3. Experimental design	Apply the fair-test protocol: identify the manipulated variable and the controls	Gatekeeping for a fair test	Group protocol
4. Construction and calibration	Connect, program, and calibrate	Technical hints: teacher escalation for hardware	Working system
5. Data collection	Take repeated measurements and graph them	Quality check of unusual values	Dataset and graph
6. CER dialogue	State claim, evidence, and reasoning	Data grounding, gatekeeping, conceptual bridging	AI-supported CER
7. Fading and transfer	Solve a new individual problem without full help	Reduced or absent prompts	Independent CER transfer

3.3. The Response-Contingency Architecture

The model's adaptivity is defined by an explicit rule set linking the quality of a pupil's response to the system's next move, so that support remains just sufficient to keep the reasoning on track (Table 8).

A general statement without data triggers a data-grounding request; data without a mechanism triggers a conceptual bridge; a scientifically inaccurate response triggers explicit correction and a teacher notification. Rule R9 sits outside the instructional sequence: if the exchange touches on a pupil's welfare, causes distress, or moves off task, the dialogue stops and the teacher is alerted.

Table 8. Response-contingency rules.

Rule	Pupil response or system state	Next move	Criterion
R1	Complete and independent	Acknowledge, advance, fade	Clear epistemic practice without a prompt
R2	Partially adequate: the prerequisite for the next step is missing	Focused prompt for the missing element	Has a claim or data, but not the full link
R3	General, without data	Data-grounding prompt	Requires specific values
R4	Data without a mechanism	Conceptual bridge	Why/how connection
R5	First unsuccessful attempt	Reformulation	Same goal, simpler language
R6	Second unsuccessful attempt	Guided choice or bounded hint	Avoiding an impasse
R7	Scientifically unsafe misconception	Explicit correction and teacher flag	Risk of consolidating an error
R8	Technically unreliable data (Table 3)	Halt interpretation, teacher check	Defective measurements are not explained
R9	Safeguarding-relevant, distressing, or off-task content	Dialogue stops; neutral holding message; teacher alerted	Welfare and safety take precedence over the task

The architecture rests on a technical assumption. Although the contingency rules are explicitly defined, they are implemented by a stochastic generative model, so the real-time classification of pupils' responses may not always be reliable.

Misrouting is possible: an adequate answer may attract an unnecessary prompt, or an unsafe one may pass. Three design decisions bound the consequences. The classification is used only for low-stakes instructional routing, never for assessment; the default action under uncertainty is the mildest one, a reformulation, with a teacher flag raised if it recurs; and every interaction is logged, so that routing accuracy can be measured and becomes part of the validation criteria in Section 3.9.

Dialogic moves are delivered as short, age-appropriate prompts, for example "Which two or three readings from your own table best support that claim?" for data grounding, "How might soil structure, drainage, or evaporation explain the different pattern?" for the conceptual bridge, and "Check my suggestion against your graph. What evidence would show that it is wrong?" for calibrated trust. Appendix A gives the full register and move specifications.

3.4. Implementation Specification of the AI-DRS System

A dialogic model implemented on a generative system is reproducible only if the configuration that produced the dialogue is stated. The reference implementation is therefore specified at the level of provider, model snapshot, prompt version, decoding settings, input format, and rule precedence.

Pupils hold no accounts and use no consumer interface. A school-controlled web application delivers the dialogue; a group signs in with a class code and a pseudonymous group label, and the application calls the provider's application programming interface server-side, under the institution's account and key. The reference configuration uses a single pinned, dated model snapshot, logged with every turn. The snapshot named here, the OpenAI API with gpt-4o-2024-11-20,

fixes the reference build rather than prescribing the model to be used; the snapshot in force at the pilot is reported verbatim, and any change triggers a re-run of the validation suite before use with pupils.

The system prompt is a fixed, version-controlled artefact (AI-DRS-SP v1.0) rather than an ad hoc instruction, and is reproduced verbatim in Appendix A. Its twelve blocks specify role and audience; the register constraint of one question per turn, at most 45 words, and no technical vocabulary not yet introduced in class; the never-answer rule, under which the system does not state the claim, the evidence, or the mechanism on the pupil's behalf and does not confirm an unjustified claim; data grounding; epistemic gatekeeping; the target-explanation reference from Table 1; the contingency rules of Table 8 and their precedence; the fading levels; the calibrated-trust statement; the refusal and safeguarding rules; and the output schema. The four response categories named in Section 3.1 are the coarse form of the eight operational classification values defined in Block 6 of Appendix A, which separate the unsupported, unreliable, and safeguarding cases that the four categories collapse together.

Decoding settings, output schema, failure fallback, and the exact input payload are specified in Table 9. Three of these are load-bearing. The 160-token limit keeps replies short enough for the target age group. A generation that fails schema validation is never shown to pupils as free text, but is retried once and then replaced by the deterministic reformulation move with a teacher flag. And nothing is transmitted beyond the group's own calibrated dataset, the pupil's current response, the phase and fade level, the two preceding turns, and the fixed target explanation, with class-list names removed by a client-side filter.

Table 9. Reference implementation specification of the AI-DRS system.

Parameter	Specification
Access architecture	School-controlled web application; no pupil accounts; class code and pseudonymous group label; server-side API calls under the institutional account.
Provider and model	A single pinned, dated snapshot, logged with every turn. The snapshot named here (OpenAI API, gpt-4o-2024-11-20) is indicative of the reference build; the snapshot current at the pilot is reported verbatim, and any change triggers re-validation.
Provider data settings	No training on submitted data; shortest available retention; only data-minimized, non-identifying content transmitted.
System prompt	AI-DRS-SP v1.0, fixed and version-controlled, reproduced verbatim in Appendix A: role, register, never-answer rule, data grounding, gatekeeping, target explanation, contingency rules and precedence, fading levels, calibrated trust, safeguarding, output schema.
Register constraints	One question per turn; maximum 45 words; no un-introduced technical vocabulary; no multi-part questions.
Decoding settings	Temperature 0.2; top-p 1.0; maximum 160 output tokens; no frequency or presence penalties; fixed seed where supported.
Output format	Schema-validated JSON with the fields classification, rule_fired, move_type, message, fade_level, and flag.
Failure fallback	One retry on schema failure, then the deterministic reformulation move with a teacher flag; generated free text is never shown unvalidated.

Continued

Input payload	Calibrated dataset (≤ 2 containers × 9 points), the pupil's current response, phase and fade level, the two preceding turns, and the fixed target explanation.
Data minimization	Client-side removal of class-list names; no pupil identifiers, no teacher notes, no prior sessions transmitted.
Logging and audit	Timestamp, snapshot, prompt version, hash of the transmitted input, classification, rule fired, move issued, and any flag.
Pre-session validation	A fixed regression suite of at least 36 synthetic pupil responses covering every category in Table 8, with no fewer than five items each for R7 and R9. Required before use with pupils: at least 80% overall, at least 70% in every category, and correct routing on every R7 and R9 item. This is a pre-session check on curated items, distinct from and weaker than the small-group usability criterion of Section 3.9.

Because more than one contingency rule can match a single response, the rules are ordered, and the system applies only the highest-priority matching rule. Table 10 sets out that hierarchy and maps every priority level onto the corresponding rule identifier in Table 8. Safeguarding overrides everything; a data-quality halt precedes any interpretive move, since interpreting unreliable measurements is worse than not interpreting them at all; scientific correction precedes gatekeeping; gatekeeping precedes data grounding and conceptual bridging; and fading is applied last, only when no higher rule has fired. When classification confidence is low, or two rules of equal priority match, the system takes the mildest available action, a reformulation, and raises a teacher flag if the difficulty recurs, so that the classifier's failure mode is conservative by design.

Table 10. Rule precedence in the AI-DRS dialogue.

Priority	Condition detected	System action	Escalation	Table 8 rule
P1 Safeguarding	Personal disclosure, distress, unsafe or off-task content.	Dialogue stops immediately; a neutral holding message is shown.	Red flag; teacher within 2 minutes.	R9
P2 Data quality	Run-level or drift flag from Table 3.	Interpretation halted; no claim, evidence, or reasoning move is issued.	Amber flag; teacher within the lesson.	R8
P3 Scientific accuracy	A scientifically unsafe misconception in the pupil response.	Brief explicit correction followed by a re-check question.	Red flag; teacher notified.	R7
P4 Gatekeeping	The prerequisite for the next step is missing.	The pupil is held at the current step with a focused prompt.	Logged only.	R2
P5 Data grounding	A general statement with no specific values.	Request for two or three specific readings from the pupil's own table.	Logged only.	R3
P6 Conceptual bridging	Data cited but no mechanism offered.	A why-or-how prompt directed at the target mechanism.	Logged only.	R4
P7 Fading	The epistemic practice is performed independently.	Acknowledge, advance, and reduce the level of prompting.	Logged only.	R1
P8 Default	Low classification confidence, or two rules of equal priority.	Reformulation in simpler language; after two failures, a bounded guided choice.	Amber flag after the second failure.	R5, then R6

3.5. Teacher Escalation, Review, and Stop Rules

Flagging is useful only if it is attached to a named person, a defined response time,

and a defined endpoint. The class teacher is the first responder and is present throughout; a live dashboard shows one row per group with the current phase, the fade level, and any open flag. A designated science lead, or a member of the research team, is the second reviewer and examines every flagged interaction within 24 hours, recording the outcome in a review log. Where a flag concerns a child's welfare rather than the science, the school's designated safeguarding lead takes over under the school's existing child-protection procedure.

Flags are graded (Table 11). A red flag pauses the dialogue automatically and alerts the teacher, who is expected to reach the group within two minutes; triggers include a safeguarding disclosure, signs of distress, content the filters should have blocked, explicit correction of a scientifically unsafe misconception, and any turn in which the system appears to have performed the pupils' reasoning. An amber flag does not pause the dialogue but requires a teacher visit within the same lesson. Routine events such as a change of fade level are logged as green entries and reviewed after the lesson.

The corrective sequence is fixed: the dialogue pauses; the teacher reads the last three turns on the dashboard, speaks with the group, and establishes whether the difficulty is scientific, technical, or affective; the teacher then resumes with corrected input, transfers the group to the static worksheet scaffold for the remainder of the session, or ends that group's AI dialogue. The second and third outcomes are recorded as fidelity deviations.

The dialogue must stop, and may not resume within that session, if the system produces scientifically incorrect content the teacher cannot correct on the spot; if it supplies the claim, evidence, or mechanism the pupils were asked to produce; if a pupil discloses safeguarding-relevant information or shows distress; if the system is used off task after one warning; if two red flags occur in one group within a session; or if a data-quality halt cannot be resolved. The group then completes the sequence with the static scaffold, so that no pupil loses the science lesson because of a system failure. The incident is logged and reported to the research team within 24 hours and, where reportable, to the ethics committee. Recurrent prompt failures identified in the weekly review trigger a prompt revision, a version increment, and a re-run of the regression suite.

Table 11. Teacher escalation protocol.

Level	Triggers	Who acts, and when	Action	Review
Red	Safeguarding disclosure; distress; blocked content that appeared; explicit correction of an unsafe misconception; the system reasoning on the pupils' behalf.	Class teacher, within 2 minutes; dialogue pauses automatically.	Read the last three turns, speak with the group, then resume, switch to the static scaffold, or stop.	Second reviewer within 24 hours; safeguarding lead where welfare is involved.
Amber	Two unsuccessful reformulations; data-quality halt; hardware fault; latency above 30 s.	Class teacher, within the same lesson; dialogue continues.	Technical or procedural fix; the affected data are marked.	Second reviewer within 24 hours.

Continued

Green	Routine events such as a change of fade level or a completed phase.	No action during the lesson.	Logged for process analysis.	Weekly review of aggregated logs.
Hard stop	Uncorrectable scientific error; the system supplies claim, evidence, or reasoning; disclosure or distress; off-task use after one warning; two red flags in one group; an unresolved data-quality halt.	Class teacher, immediately.	AI dialogue ends for that group; the sequence is completed with the static scaffold.	Reported to the research team within 24 hours, and to the ethics committee where reportable; prompt revision and re-validation follow.

3.6. Fading, Uptake and Engagement

The presence of a scaffold does not guarantee uptake: a conceptual bridge may lead to full articulation, simple acceptance, or no response at all. The model therefore separates the system's moves from the quality of the pupil's uptake and treats engagement as a function of the substance of the response, data use, revision, and persistence rather than the number of turns. Fading occurs when the pupil performs the required practice independently and proceeds through four levels: a direct prompt naming the step, a delayed prompt issued only after a turn's pause, a minimal prompt that acknowledges the response without asking anything further, and no prompt at all. Support is reduced by one level at a time and may return temporarily if performance declines. Block 9 of Appendix A states the corresponding fade_level values.

3.7. Assessment Instruments

Two instruments make reasoning visible. The CER rubric (Table 12) scores six dimensions to a maximum of twelve points. Its Mechanism dimension is scored against the target mechanism of Table 1 rather than a general standard: a score of 2, partial mechanism, requires a correct but incomplete appeal to particle size, pore space, or organic matter, for example naming faster drainage without relating it to soil structure, whereas a score of 3, full and consistent, requires the pupil to link a stated numerical difference in the rate of decline to the particle-size and pore-space account of Table 1 while making no claim outside the bounds listed there. Raters use the same Table 1 reference that is supplied to the AI-DRS system, so that the human and machine standards coincide. The chat-log coding scheme classifies the system's moves and the pupil's responses so that only pupil-expressed synthesis counts as independent reasoning.

Table 12. Claim-Evidence-Reasoning (CER) rubric.

Dimension	0	1	2	3
Claim	Absent/incorrect	Vague	Clear and testable	-
Evidence relevance	No reference	Partially relevant	Relevant and sufficient	-
Data specificity	No data	General trend	Specific values/changes	-
Evidence-claim link	Absent	Implicit connection	Explicit connection	-

Continued

Mechanism (against Table 1)	Absent/wrong	Descriptive	Partial mechanism: correct but incomplete appeal to particle size, pore space, or organic matter	Full and consistent: numerical rate difference linked to the particle-size and pore-space account of Table 1
Limitations	Not recognized	Recognizes a limit/alternative	-	-

3.8. TPACK Alignment

The model is designed so that technology is not treated separately from the science content and the pedagogical goals it supports (Mishra et al., 2023). This view is consistent with computational-pedagogy frameworks for STEAM education, which emphasize the coordinated interaction of content, pedagogical practice, computational methods, and technological tools (Psycharis et al., 2020), and with work integrating AI-based assessment models and authentic STEAM engineering tasks in contexts such as precision agriculture (Xenakis, 2025). Each phase involves a specific combination of content, pedagogical, and technological knowledge (Table 13); the analysis and explanation phases represent the fullest TPACK intersection, where data logging and the AI-DRS dialogue work with CER scaffolding to promote mechanistic reasoning.

Table 13. TPACK alignment of the intervention phases.

Phase	Content knowledge	Pedagogical knowledge	Technology	Intersection
1. Orientation	Moisture, plants, change	Elicitation	-	PCK
2. Prediction	Variables and mechanisms	Inquiry learning	-	PCK
3. Experimental design	Fair testing	Collaborative inquiry	micro:bit design	TPACK
4. Construction and calibration	Measurement and calibration	Learning by making	Nezha, sensor, MakeCode	TCK/TPK
5. Data collection	Patterns and graphs	CER scaffolding	Data logging	TPACK
6. CER dialogue	Evidence-mechanism	Dialogic scaffolding	AI-DRS	TPACK
7. Fading and transfer	Applying the model	Fading/metacognition	Reduced AI support	TPACK

3.9. Readiness for Validation

Finally, the development strand sets out the criteria that must be met before the pilot can proceed. A panel of five to seven experts in science education, primary teaching methods, educational technology, and research ethics will rate each element for relevance, clarity, feasibility, developmental appropriateness, scientific accuracy, and safeguarding, and will record item- and scale-level content-validity indices (I-CVI ≥ .78 and S-CVI/Ave ≥ .90).

The minimum requirements are: no unresolved objection to scientific accuracy; a successful review of the prompts against misconception, privacy, and refusal cases; routing accuracy of at least 80% in the small-group usability test, measured

as the percentage of authentic pupil responses that the system categorizes as an independent human coder does, with no category of **Table 8** below 70% and with correct routing on every safeguarding (R9) and scientific-accuracy (R7) item, since an aggregate threshold can be satisfied while the two safety-critical rules fail; the pre-session regression suite of **Table 9** is scored against the same per-category floors but is a separate and weaker check, because its items are curated and easier than authentic responses, and meeting it does not discharge the usability criterion; stable hardware operation and logging in at least 90% of trials; pupil comprehension of instructions without extended explanation; teacher feasibility within the timetabled slots; a documented dry run of the escalation protocol including at least one simulated red flag; and a pilot scoring agreement of $\kappa \geq .70$.

4. Discussion

The model can be situated against three areas of prior research. First, GenAI in science often assumes the role of a single source of knowledge (Cooper, 2023), which, combined with young children's tendency to over-trust conversational agents (Movahed & Martin, 2025), explains why adding an answer-provider role to reasoning instruction would backfire. AI-DRS reverses that role: the technology is designed to ask rather than to answer, and its data-grounding and gatekeeping rules keep pupils' own measurements at the center of every explanation. It thereby extends the finding that a customized chatbot can support reasoning and argumentation in secondary science (Tang & Putra, 2026) to primary education, adding pupil-generated sensor data, explicit gatekeeping, and structured fading. This positioning is consistent with earlier evidence that digital technologies in young children's science learning are most productive when they complement hands-on inquiry and teacher mediation rather than replace them (Kalogiannakis et al., 2018).

Second, the model sharpens the literature that reads AI through a scaffolding lens. Reviews there discuss scaffolding as personalized learning paths, tailored support, and immediate feedback, but are largely based in higher education and tend to treat scaffolding as adaptive delivery (Bai et al., 2026; Cai et al., 2025; Wan Hamedi et al., 2025). Chatbot effects appear weaker for younger learners and to diminish over time (Wu & Yu, 2024), and the benefits of GenAI are entangled with risks to cognition and agency (Wu & Yu, 2024; Yan et al., 2025). This is not an argument against GenAI in primary education; it is what makes the model's developmental and safety features load-bearing, since attunement, fading, and teacher involvement are necessary conditions at this age rather than optional extras. The model also foregrounds a dimension that delivery-centered accounts underplay: the dialogic and argumentative quality of the interaction (Alexander, 2018; Muhonen et al., 2016; Vrikki et al., 2019; Wegerif & Casebourne, 2026).

Third, the model must navigate tensions rather than remove them. The first concerns epistemic agency: heavy reliance on GenAI is associated with reduced critical thinking and weaker self-regulation (Lee et al., 2025; Yan et al., 2025). AI-

DRS responds with fading criteria, a stance that withholds answers, and assessment only of knowledge pupils express themselves, though whether these safeguards work remains open. The second concerns trust and misinformation: young children are credulous users, so a system that occasionally errs could instill misconceptions, which is why teacher verification and AI literacy are built in from the outset (Movahed & Martin, 2025; Ng et al., 2024). The third concerns the teacher, since the model assumes the technological, pedagogical, and content knowledge that makes professional development a precondition (Mishra et al., 2023). The fourth concerns equity, since unequal access to devices and connectivity could widen gaps unless inclusion is addressed explicitly.

The main theoretical contribution of this article is a sharper account of scaffolding when the scaffold is generative and conversational. Classically, a scaffold is defined by contingency, fading, and the transfer of responsibility to the learner (van de Pol et al., 2010; Wood et al., 1976). A generative system satisfies the first well, assessing a response and adjusting its next move, but it does not deliver the other two: it has no intrinsic mechanism for reducing support, and children who trust these systems are unlikely to question or refuse it (Movahed & Martin, 2025). Left unregulated, the result is a scaffold that never diminishes, which in classical terms is a permanent aid rather than a scaffold. Better prompts cannot resolve this; it reflects a difference in kind between human scaffolding, which rests on pedagogical judgment, and automated scaffolding, whose responses are generated probabilistically within predefined rules.

AI-DRS builds the consequences of that difference into its design. If fading will not emerge from the technology, it must be written into the interaction, with explicit criteria for when support is reduced and who decides. Success must therefore be measured by the independent explanation that follows the conversation, not by the completed exchange. This is why the model credits only pupil-generated synthesis, treats the AI-free transfer task as the key outcome, and specifies fading as a requirement rather than a hope. Its value is greatest where a teacher cannot supply tailored questioning to every group: the system increases the available dialogue while the teacher retains responsibility for scientific accuracy, safety, and the decision to use it at all.

5. Conclusions

This study developed the AI-DRS Physical Inquiry Model as a structured, safety-oriented approach to supporting scientific reasoning in primary education. The model connects pupil-generated sensor data, contingent dialogue, CER scaffolding, and systematic fading within a Nezha-micro:bit soil-moisture investigation. Its central premise is that GenAI has educational value only when it helps pupils turn measurements into evidence and evidence into mechanistic explanations, while progressively transferring responsibility back to the learner.

The study integrates elements previously examined largely in isolation: pupil-generated sensor data, dialogic GenAI support, CER scaffolding, and the system-

atic reduction of technological assistance. Physical-computing activities let pupils generate authentic measurements but rarely support the move from measurement to evidence and mechanism (Przybylla & Romeike, 2014; Kalelioğlu & Sentance, 2020; Lin et al., 2023), whereas GenAI systems sustain personalized dialogue but may detach from learners' observations or reason on their behalf (Cooper, 2023; Yan et al., 2024). AI-DRS addresses this dual gap by keeping every substantive dialogue move grounded in pupils' own measurements and positioning GenAI as a temporary reasoning scaffold rather than an answer-generating authority. Its first contribution is therefore a data-grounded bridge between physical inquiry and dialogic artificial intelligence in primary science.

A second contribution is the explicit response-contingency and fading architecture. Rather than a general prompt or an unrestricted chatbot conversation, AI-DRS specifies how the system responds to complete, partial, unsupported, scientifically inaccurate, and technically unreliable reasoning; introduces gatekeeping criteria that block progress until defined requirements are met; and states when support is reformulated, intensified, reduced, temporarily restored, or transferred to the teacher. Fading thereby becomes an operational component of the design rather than an assumed by-product of repeated AI use, extending conventional accounts of scaffolding into a generative conversational environment (Wood et al., 1976; van de Pol et al., 2010).

A third contribution concerns how learning and learner agency are evaluated. The model distinguishes the dialogic move produced by the AI from the pupil's uptake of it, avoiding the assumption that exposure to an AI-generated explanation is equivalent to learning. Only explanations pupils articulate themselves count as evidence of reasoning, and the central outcome is an independent CER explanation produced after AI support has been removed. With teacher oversight, calibrated-trust prompts, age-appropriate safeguarding, and mediated access through a school-controlled environment, this emphasis on post-scaffold transfer offers a framework for addressing over-trust, cognitive dependence, and epistemic agency in GenAI-supported learning (Movahed & Martin, 2025; Yan et al., 2025).

The contribution is an intervention ready for validation rather than proof of long-term effectiveness, and its limitations follow. The model has been designed but not tested in authentic classrooms; the real-time classification performed by a stochastic generative model may be unreliable; a pilot cannot demonstrate long-term effectiveness; the number of randomized groups is small, so the pilot is designed to bound design parameters such as the intraclass correlation and the design effect rather than estimate them precisely or test efficacy; both conditions are taught in the same room within the same session, so contamination between conditions cannot be excluded and would attenuate any observed difference; and novelty, group-work effects, model variability, sensor error, and teacher influence may all complicate results. The next steps are expert validation and the pilot with an active control condition, assessing feasibility and the proposed mechanism: whether data-grounded dialogue helps pupils construct their own scientific expla-

nations without AI. If it does, GenAI should function not as an answer provider but as a temporary dialogic scaffold that elicits, probes, and progressively transfers reasoning responsibility to the learner.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Alexander, R. (2018). Developing Dialogic Teaching: Genesis, Process, Trial. *Research Papers in Education, 33,* 561-598. https://doi.org/10.1080/02671522.2018.1481140

Anthropic (2025). *Consumer Terms of Service*. https://www.anthropic.com/legal/terms

Bai, S. T., Yeung, S. S., & Lo, C. K. (2026). Enhancing the Effect of AI-Assisted Learning: The Use of Scaffolding Strategies to Develop Students' Prompt Engineering Skills. *Interactive Learning Environments,* 1-22. https://doi.org/10.1080/10494820.2026.2649547

Bell, R. M., & McCaffrey, D. F. (2002). Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples. *Survey Methodology, 28,* 169-181.

Cai, L., Msafiri, M. M., & Kangwa, D. (2025). Exploring the Impact of Integrating AI Tools in Higher Education Using the Zone of Proximal Development. *Education and Information Technologies, 30,* 7191-7264. https://doi.org/10.1007/s10639-024-13112-0

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics, 90,* 414-427. https://doi.org/10.1162/rest.90.3.414

Cao, J., Chan, S. W. T., Garbett, D. L., Denny, P., Nassani, A., Scholl, P. M. et al. (2021). Sensor-Based Interactive Worksheets to Support Guided Scientific Inquiry. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference* (pp. 1-7). ACM. https://doi.org/10.1145/3459990.3460716

Cooper, G. (2023). Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *Journal of Science Education and Technology, 32,* 444-452. https://doi.org/10.1007/s10956-023-10039-y

Diola, W. Y., Jalon Jr., J. B., & Prudente, M. S. (2025). Exploring the Use of Claim-Evidence-Reasoning in Promoting Scientific Reasoning Skills of Elementary School Students. *Anatolian Journal of Education, 10,* 203-214. https://doi.org/10.29333/aje.2025.10115a

ELECFREAKS (n.d.). *Nezha Inventor's Kit for Micro:Bit: Product Documentation and Learning Cases*. https://wiki.elecfreaks.com/en/microbit/building-blocks/nezha-inventors-kit/

Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics, 32,* 341-370. https://doi.org/10.3102/1076998606298043

Kalelioğlu, F., & Sentance, S. (2020). Teaching with Physical Computing in School: The Case of the Micro:Bit. *Education and Information Technologies, 25,* 2577-2603. https://doi.org/10.1007/s10639-019-10080-8

Kalogiannakis, M., Ampartzaki, M., Papadakis, S., & Skaraki, E. (2018). Teaching Natural Science Concepts to Young Children with Mobile Devices and Hands-On Activities. A Case Study. *International Journal of Teaching and Case Studies, 9,* 171-183. https://doi.org/10.1504/ijtcs.2018.090965

Kalogiannakis, M., Papakonstantinou, N., & Sotiropoulos, D. (2025). From Support Tool to Learning Partner: A Systematic Review of GenAI Integration in University Science

Labs. *Creative Education, 16,* 1364-1401. https://doi.org/10.4236/ce.2025.169083

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R. et al. (2025). The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers. In Association for Computing Machinery (Ed.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-22). ACM. https://doi.org/10.1145/3706598.3713778

Lin, X.-F., Hwang, G.-J., Wang, J., Zhou, Y., Li, W., Liu, J., & Liang, Z.-M. (2023). Effects of a Contextualised Reflective Mechanism-Based Augmented Reality Learning Model on Students' Scientific Inquiry Learning Performances, Behavioural Patterns, and Higher Order Thinking. *Interactive Learning Environments, 31,* 6931-6951. https://doi.org/10.1080/10494820.2022.2057546

Masters, H., & Docktor, J. (2022). Preservice Teachers' Abilities and Confidence with Constructing Scientific Explanations as Scaffolds Are Faded in a Physics Course for Educators. *Journal of Science Teacher Education, 33,* 786-813. https://doi.org/10.1080/1046560x.2021.2004641

McNeill, K. L., & Krajcik, J. S. (2012). *Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing*. Pearson.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting Students' Construction of Scientific Explanations by Fading Scaffolds in Instructional Materials. *Journal of the Learning Sciences, 15,* 153-191. https://doi.org/10.1207/s15327809jls1502_1

McNeish, D. M., & Stapleton, L. M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review, 28,* 295-314. https://doi.org/10.1007/s10648-014-9287-x

Mishra, P., Warr, M., & Islam, R. (2023). TPACK in the Age of ChatGPT and Generative AI. *Journal of Digital Learning in Teacher Education, 39,* 235-251. https://doi.org/10.1080/21532974.2023.2247480

Movahed, S. V., & Martin, F. G. (2025). Ask Me Anything: Exploring Children's Attitudes toward an Age-Tailored AI-Powered Chatbot. *International Journal of Artificial Intelligence in Education, 35,* 3979-4001. https://doi.org/10.1007/s40593-025-00523-4

Muhonen, H., Rasku-Puttonen, H., Pakarinen, E., Poikkeus, A., & Lerkkanen, M. (2016). Scaffolding through Dialogic Teaching in Early School Classrooms. *Teaching and Teacher Education, 55,* 143-154. https://doi.org/10.1016/j.tate.2016.01.007

Ng, D. T. K., Su, J., Leung, J. K. L., & Chu, S. K. W. (2024). Artificial Intelligence (AI) Literacy Education in Secondary Schools: A Review. *Interactive Learning Environments, 32,* 6204-6224. https://doi.org/10.1080/10494820.2023.2255228

Ntourou, V., Kalogiannakis, M., & Psycharis, S. (2021). A Study of the Impact of Arduino and Visual Programming in Self-Efficacy, Motivation, Computational Thinking and 5th Grade Students' Perceptions on Electricity. *Eurasia Journal of Mathematics, Science and Technology Education, 17,* em1960. https://doi.org/10.29333/ejmste/10842

OpenAI (2025). *Terms of Use*. https://openai.com/policies/row-terms-of-use/

Petousi, V., & Sifaki, E. (2020). Contextualising Harm in the Framework of Research Misconduct. Findings from Discourse Analysis of Scientific Publications. *International Journal of Sustainable Development, 23,* 149-174. https://doi.org/10.1504/ijsd.2020.115206

Przybylla, M., & Romeike, R. (2014). Physical Computing and Its Scope—Towards a Constructionist Computer Science Curriculum with Physical Computing. *Informatics in Education, 13,* 225-240. https://doi.org/10.15388/infedu.2014.14

Psycharis, S., Kalovrektis, K., & Xenakis, A. (2020). A Conceptual Framework for Compu-

tational Pedagogy in STEAM Education: Determinants and Perspectives. *Hellenic Journal of STEM Education, 1,* 17-32. https://doi.org/10.51724/hjstemed.v1i1.4

Sentance, S., Waite, J., Hodges, S., MacLeod, E., & Yeomans, L. (2017). "Creating Cool Stuff": Pupils' Experience of the BBC Micro:Bit. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (pp. 531-536). Association for Computing Machinery. https://doi.org/10.1145/3017680.3017749

Sotiropoulos, D., Xenakis, A., Kalogiannakis, M., & Taşar, M. F. (2026). The Interactive Design Process Framework (IDPF): Utilizing GenAI as a Collaborative Agent for Creating STEAM Projects. *Hellenic Journal of STEM Education, 5,* 1-18. https://doi.org/10.51724/hjstemed.v5i1.76

Spasopoulos, T., Sotiropoulos, D., & Kalogiannakis, M. (2025). Generative AI in Pre-Service Science Teacher Education: A Systematic Review. *Advances in Mobile Learning Educational Research, 5,* 1501-1523. https://doi.org/10.25082/amler.2025.02.007

Tang, K.-S., & Putra, G. B. S. (2026). Generative AI as a Dialogic Partner: Enhancing Multiple Perspectives, Reasoning, and Argumentation in Science Education with Customized Chatbots. *Journal of Science Education and Technology, 35,* 128-140. https://doi.org/10.1007/s10956-025-10240-1

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in Teacher-Student Interaction: A Decade of Research. *Educational Psychology Review, 22,* 271-296. https://doi.org/10.1007/s10648-010-9127-6

Vrikki, M., Wheatley, L., Howe, C., Hennessy, S., & Mercer, N. (2019). Dialogic Practices in Primary School Classrooms. *Language and Education, 33,* 85-100. https://doi.org/10.1080/09500782.2018.1509988

Wan Hamedi, W. H., Awang Ali, F. D., Abdullah, W. Y., Ab Hamid, H., Mohammad Shuhaimi, N. I., & Mohamad Amir, M. (2025). AI as a Digital Scaffold: An Integrative Review of Vygotsky's Zone of Proximal Development in Modern Education. *International Journal of Modern Education, 7,* 579-589. https://doi.org/10.35631/ijmoe.726038

Wegerif, R., & Casebourne, I. (2026). A Dialogic Theoretical Foundation for Integrating Generative AI into Pedagogical Design. *British Journal of Educational Technology, 57,* 639-654. https://doi.org/10.1111/bjet.70026

Wood, D., Bruner, J. S., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry, 17,* 89-100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Wu, R., & Yu, Z. (2024). Do AI Chatbots Improve Students Learning Outcomes? Evidence from a Meta-Analysis. *British Journal of Educational Technology, 55,* 10-33. https://doi.org/10.1111/bjet.13334

Xenakis, A. (2025). AI-Based Models for Assessing STEAM Engineering Literacy for University Students: The Case of Digital Systems for Precision Agriculture. *Hellenic Journal of STEM Education, 4,* 1-9. https://doi.org/10.51724/hjstemed.v4i1.37

Xenakis, A., Dimos, I., Feidakis, M., Sotiropoulos, D., Kalovrektis, K., & Nikolaou, G. (2025). An LLM-Based Smart Repository Platform to Support Educators with Computational Thinking, AI, and STEM Activities. In S. Papadakis, & M. Kalogiannakis (Eds.), *Empowering STEM Educators with Digital Tools* (pp. 107-136). IGI Global. https://doi.org/10.4018/979-8-3693-9806-7.ch005

Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and Challenges of Generative Artificial Intelligence for Human Learning. *Nature Human Behaviour, 8,* 1839-1850. https://doi.org/10.1038/s41562-024-02004-5

Yan, L., Pammer-Schindler, V., Mills, C., Nguyen, A., & Gašević, D. (2025). Beyond Effi-

ciency: Empirical Insights on Generative AI's Impact on Cognition, Metacognition and Epistemic Agency in Learning. *British Journal of Educational Technology, 56,* 1675-1685. https://doi.org/10.1111/bjet.70000

Appendix A: AI-DRS-SP v1.0: System Prompt, Verbatim

The text below is the complete system prompt used in the reference implementation described in Section 3.4. It is reproduced without modification so that the dialogue behavior reported in any subsequent empirical study can be replicated and audited. The bracketed headings mark the twelve blocks and are part of the prompt; the target explanation and the dataset named in Blocks 4 and 7 are inserted at run time from the group's own record.

[BLOCK 1 - ROLE AND AUDIENCE]

You are a science-inquiry partner for pupils aged 10 to 12 working in a small group in a primary school classroom. You are not a teacher, not an assessor, and not an encyclopedia. Your only purpose is to help this group turn their own soil-moisture measurements into a claim, evidence, and reasoning. A human teacher is present in the room at all times and can see every turn of this conversation.

[BLOCK 2 - REGISTER]

Write one question per turn. Never more than one. Use no more than 45 words. Use short sentences and everyday words. Do not use any technical term that does not already appear in the pupils' worksheet or in the target explanation given in Block 7. Never ask a multi-part question. Never use bullet points, headings, or emoji.

[BLOCK 3 - NEVER-ANSWER RULE]

You must never state the claim, the evidence, or the reasoning on the pupils' behalf, in whole or in part, even if the pupils ask you to, even if they say they are stuck, and even if they have already given a nearly complete answer. You must never confirm a claim that the pupils have not yet justified with their own data. You may not supply the mechanism. You may not summarize the pupils' explanation back to them as if it were finished. If the pupils ask you directly for the answer, reply with a question that returns them to their own data.

[BLOCK 4 - DATA GROUNDING]

Every substantive move you make must refer to the group's own dataset, which is supplied to you in the input payload. Do not accept a general statement such as "it dried out more" or "sand is drier". Require at least two specific readings, with their times and containers, before you allow the pupils to move from evidence to reasoning. If the pupils cite a reading that does not appear in their dataset, ask them to check their table; do not correct the number for them.

[BLOCK 5 - EPISTEMIC GATEKEEPING]

Hold the pupils at the current step until its requirement is met.

Evidence claim: the pupils must have stated a claim that is testable with their dataset and that names the comparison, not just an outcome.

Evidence to reasoning: the pupils must have cited at least two specific readings from each container that are relevant to the claim.

Reasoning to completion: the pupils must have connected the stated difference in the rate of decline to a physical mechanism in their own words.

Do not advance the phase yourself. Advancing is signaled by the application, not by you.

[BLOCK 6 - CLASSIFICATION]

Classify the pupils' most recent response as exactly one of: complete_independent, partially_adequate, general_no_data, data_no_mechanism, unsuccessful_attempt, un-

safe_misconception, unreliable_data, safeguarding_or_off_task. This classification is used only to choose your next instructional move. It is never an assessment of the pupils and is never shown to them. If you are not confident, classify as unsuccessful_attempt and reformulate; the application raises an amber flag after a second consecutive unsuccessful attempt, as set out in Block 8, P8.

[BLOCK 7 - TARGET EXPLANATION]

The target explanation for this investigation is fixed and is supplied in the input payload as target_explanation. You must not accept an explanation that contradicts it, and you must not extend the pupils' claim beyond the bounds it states. The current target is: sandy soil has larger particles and larger pore spaces and little organic matter, so water drains through it faster and more of it is exposed to the air; the finer particles and the organic matter in compost hold water against drainage and evaporation. Out-of-scope claims, which you must not endorse or encourage, are: which soil is better for plants; effects on plant growth or health; the index as an absolute amount of water; behavior beyond 40 minutes or after re-watering; and causal claims about light, temperature, or container size.

[BLOCK 8 - CONTINGENCY RULES AND PRECEDENCE]

Exactly one rule fires per turn. Evaluate in this order and stop at the first match.

P1 safeguarding (R9): if the response contains a personal disclosure, signs of distress, or unsafe or off-task content, stop the dialogue, return the neutral holding message "Thank you. Let's pause here and ask your teacher to come over.", set flag to "red", and issue no further question.

P2 data quality (R8): if the payload marks the run as drift-flagged or run-level-flagged, halt interpretation, issue no claim, evidence, or reasoning move, set flag to "amber", and tell the pupils that the teacher will check the equipment.

P3 scientific accuracy (R7): if the response contains a scientifically unsafe misconception, give one brief explicit correction of the misconception only, follow it with one re-check question, and set flag to "red".

P4 gatekeeping (R2): if the prerequisite for the next step is missing, hold the pupils at the current step with one focused prompt for the missing element.

P5 data grounding (R3): if the response is general with no specific values, ask for two or three specific readings from the pupils' own table.

P6 conceptual bridging (R4): if data are cited but no mechanism is offered, ask one why-or-how question directed at the target mechanism, without naming the mechanism.

P7 fading (R1): if the epistemic practice was performed independently, acknowledge briefly, and reduce the prompt level by one.

P8 default (R5, then R6): if classification confidence is low, or two rules of equal priority match, reformulate the previous question in simpler language with the same goal. On a second consecutive unsuccessful attempt, offer a bounded guided choice of two options, neither of which states the mechanism, and set flag to "amber".

[BLOCK 9 - FADING LEVELS]

fade_level 3, direct prompt: ask the full question, naming the step.

fade_level 2, delayed prompt: wait one turn; ask only "What would you add to that?".

fade_level 1, minimal prompt: acknowledge only; ask nothing.

fade_level 0, no prompt: return an empty message.

Reduce fade_level by one when rule R1 fires. Increase it by one, to a maximum of 3, if the pupils' next response falls below the level of the previous one. Never reduce fade_level by more than one step in a turn.

[BLOCK 10 - UNCERTAINTY AND CALIBRATED TRUST]

You can be wrong. At least once per phase, ask the pupils to test what you have said against their own graph, for example: "Check my suggestion against your graph. What evidence would show that it is wrong?" Never claim certainty about the pupils' data. If you cannot tell what the data show, say so and ask the pupils to look.

[BLOCK 11 - REFUSAL AND SAFEGUARDING]

Do not discuss anything other than this investigation. Do not answer questions about yourself, about other pupils, about health, medicine, family, or personal circumstances, or about topics outside the lesson. Do not give advice of any kind. If any such content appears, apply P1 immediately: stop, return the holding message, set flag to "red", and end the exchange. Do not repeat back any name, address, school, or other identifying detail that appears in a pupil's text.

[BLOCK 12 - OUTPUT FORMAT]

Return a single JSON object and nothing else, conforming to this schema:

{"classification": string, "rule_fired": string, "move_type": string, "message": string, "fade_level": integer, "flag": string}

classification is one of the eight values in Block 6. rule_fired is one of R1-R9. move_type is one of acknowledge, focused_prompt, data_grounding, conceptual_bridge, reformulation, guided_choice, correction, halt, stop. message is the text shown to the pupils and obeys Block 2. fade_level is 0 to 3. flag is one of green, amber, red. Return nothing outside this object. If the object you return fails schema validation, the application discards it and substitutes the deterministic fallback move server-side; no unvalidated generation is ever shown to a pupil.