Cognitive Seeds as an Integrative Design Framework for Human Understanding in AI-Mediated Learning

John Gao

Humanity to Eternity (H2E)

Independent Researcher / Convener, H2E

United States

October 2026

Crossref DOI: Pending Registration

Abstract

Generative artificial intelligence can improve an assisted answer without improving the understanding its user can subsequently exercise. This conceptual article presents Cognitive Seeds as an integrative design framework, not a new explanatory theory of cognition. A cognitive seed denotes an initial, revisable relational understanding whose proposed generative property requires evidence of further questioning, application, and learning from new information. The framework distinguishes external carriers, a selected minimal knowledge set, and a learner-developed minimal cognitive core. A micro-worldview is an unmeasured pedagogical heuristic for connecting aspects of that core across self, others, action, and uncertainty. Comparison with spiral curricula, threshold concepts, big ideas, and conceptual-change approaches locates the contribution in coordinated design commitments rather than novel psychological mechanisms. These commitments include explicit relation-to-carrier maps, context-bound omission tests, and assessment of independent use and preparation for future learning. Seven activity descriptions support observation without constituting a validated scale. A staged research agenda separates feasibility from effectiveness, controls prompt–assessment alignment, specifies the AI's instructional role, and distinguishes justified revision from model deference. No participant data or efficacy claims are presented. The framework earns additional value only if these commitments improve design decisions or outcomes beyond credible existing approaches.

Keywords: cognitive seeds; instructional design; generative artificial intelligence; relational understanding; preparation for future learning; transfer

1. Introduction

An educational answer has at least two possible achievements: it can satisfy a task requirement, and it can contribute to the learner's subsequent ability to understand and act. These achievements may coincide, but their relationship requires examination. When generative AI supplies explanations, arguments, or completed solutions, evaluating the

resulting artifact alone cannot establish which parts of the understanding remain available to the person who used it.

Learning and performance can diverge (Soderstrom & Bjork, 2015). In high-school mathematics, Bastani et al. (2025) compared ordinary instruction with GPT Base, a general-purpose chat interface, and GPT Tutor, which used teacher-informed instructions and materials to guide assistance. Assisted practice improved in both AI conditions. On a subsequent examination without AI, GPT Base performed worse than the control; GPT Tutor did not show that disadvantage. This is evidence about those instructional conditions, not a general finding that generative AI reduces cognition or that safeguards guarantee durable learning.

Cognitive Seeds asks what initial structures of understanding help a learner question, interpret, and revise externally supplied material. The author traces the proposal to an educational work written as Si Jiu, referred to here by the published manuscript "New Education: Cultivating Seeds of Cognition (English only): ChatGPT Era " (Gao, 2024). The curriculum uses concise poetry and progressively elaborated activities as starting materials. This article develops a conceptual framework and an evaluation agenda; it does not report an evaluation of that curriculum.

The author also describes a published manuscript, "A Guide for Planting the Future of Education: Cognitive seeds textbook toolkit" (Gao, 2025), containing instructional modules and lesson examples.

The contribution is an integrative design framework, expressed through a practical vocabulary. It coordinates three commitments: identify relations separately from their carriers; test a starting set's adequacy and dispensability for a specified learner group; and examine independent application, revision, and subsequent learning rather than assisted output alone. Its components have antecedents. The claim is that making their coordination explicit may improve instructional specification and evaluation, not that a new cognitive mechanism has been discovered. Childhood cultivation and adult AI interaction remain separate applications requiring separate evidence.

2. Conceptual approach and scope

The article uses conceptual analysis and design-oriented synthesis rather than a systematic review. Sources were selected for relevance to knowledge organization, learner-generated elaboration, assistance, transfer, and instructional design. The argument compares the proposed terms with established approaches, identifies counterexamples, and specifies tests of design commitments. This selective process cannot establish that the literature lacks another framework combining similar commitments; novelty is consequently limited to the particular integration and its explicit evaluation requirements.

Three kinds of claim are distinguished throughout. Descriptive claims summarize cited research. Construct definitions specify how the proposal uses its terms. Design propositions state expectations that remain to be tested. Examples illustrate proposed relationships; they are not observations of participants. No new human-subject data are reported.

The term cognitive system is used at a functional and educational level. It does not identify a single neural structure, establish a theory of consciousness, or imply that all cognitive functions have one common mechanism. Likewise, the proposed seven activity dimensions are a working vocabulary for task design. They are not asserted to be exhaustive, independent psychological faculties or a validated measurement scale.

3. The cognitive system and its interaction with the world

3.1 A functional definition

A personal cognitive system is provisionally defined here as the evolving organization of perceptual, mnemonic, conceptual, affective, and goal-directed processes through which a person interprets experience, forms judgments, and guides action. The definition emphasizes relationships among processes. It allows the same information to have different significance for people with different experiences, concerns, and available knowledge.

For educational purposes, cognitive development can be examined through recurring interactions among experience, interpretation, action, feedback, and revision. A learner observes an event, relates it to prior understanding, proposes an explanation, acts or seeks further information, and reconsiders the explanation in light of consequences. This is a heuristic sequence, not a claim that cognition always proceeds in discrete stages. Some responses are implicit, affective, habitual, or simultaneous, and social interaction can alter every part of the sequence.

Cultivation targets conditions that educators can influence within this recurring process: which features receive attention, what vocabulary and examples support interpretation, whether action produces interpretable feedback, and whether learners have a meaningful opportunity to revise. Instruction should specify those opportunities rather than infer development from exposure to the carrier alone.

A child rebuilding a fallen block tower illustrates the distinction between an event and its interpretation. The initial explanation might attribute collapse to an accidental touch. Further attempts might reveal the importance of base width. The educational question is whether the learner can explain why changing the base matters and recognize comparable instability in a new construction. Successful reconstruction by copying an adult is a different observation.

3.2 Knowledge organization and external support

The importance of organized knowledge is compatible with schema-based approaches to learning. Sweller (1988) links problem-solving demands and schema acquisition, warning that activity can consume processing resources without producing the desired learning. A compact curriculum therefore cannot justify itself by giving learners less information. It must still supply the knowledge needed for the intended reasoning, and its tasks must be manageable for the particular learner.

An account centered on the learner need not treat external resources as threats. Clark and Chalmers (1998) argue that environmental resources can participate in cognitive processes, while Risko and Gilbert (2016) examine cognitive offloading and its

metacognitive regulation. These perspectives challenge an equation between cognitive agency and performing every operation internally. The concern of Cognitive Seeds is narrower: whether learners can understand the purpose and limits of assistance, evaluate relevant claims, and retain capabilities required by the educational goal.

Accordingly, the framework distinguishes a temporary dependence appropriate to learning from an unexplained loss of opportunity to practice the target activity. A worked example, a teacher's explanation, and an AI-generated suggestion may each support learning. Their educational value depends partly on what the learner does with them. Removing assistance can help examine a specified capability; it is not a universal test of human worth or cognitive independence.

4. Cognitive seeds and a minimal cognitive core

4.1 The seed as a relational understanding

A cognitive seed is defined as an initial, revisable relational understanding that can support further explanation, questioning, and application. The seed is a learner's understanding, whereas a poem, story, diagram, or experience is a possible carrier. Remembering a carrier does not establish that its proposed relations have been understood.

For example, a story can carry a relation between actions and consequences. Recalling the story is evidence of recall; explaining the relation and its limits in an unfamiliar case supplies different evidence. The proposed generative property goes further: the learner should be able to formulate a relevant new question, identify information that could resolve it, and use new evidence to extend or revise an explanation. Section 8.4 specifies a preparation-for-future-learning task. No single successful answer establishes a seed as a distinct internal entity.

Four requirements constrain the proposal. The relation must be understandable at the intended level of development. It must connect with more than one possible experience or question. It must permit elaboration beyond reproduction of the original material. Finally, it must remain open to correction. These are proposed selection and evaluation criteria, not established necessary and sufficient conditions for learning.

Operational support for generativity should require more than a novel phrase. A relevant question must address an uncertainty in the case; an information choice must distinguish plausible alternatives; and an extension must use newly supplied evidence appropriately. Independent assessors should accept more than one defensible answer and check whether the same behavior occurs after replacing the carrier. Rote question templates, irrelevant novelty, and confident repetition do not meet these criteria.

A concept can provide an external carrier when presented as a term, rule, or worked example. The seed designation concerns the learner's initial, revisable use of the relation, not the concept's label or an assumed stage of mastery. More developed conceptual understanding can retain the same expandable and revisable properties; seed and mature understanding need not be sharply separated psychological kinds.

4.2 What minimal means

A minimal knowledge set is an externally selected collection of materials judged sufficient to begin a specified learning pathway. A minimal cognitive core is the learner's developing organization of relations that makes those materials usable. The two cannot be assumed to coincide: educators can deliver the same materials while learners construct different understandings.

Minimality is relative to a goal, population, and context. It does not mean the smallest universal list of facts, nor does it imply that foundational knowledge can replace subject expertise. Selection involves a trade-off between accessibility and adequate explanatory resources. A defensible starting set should be compact enough to revisit but rich enough to distinguish relevant cases, support reasons, and expose its own limits.

This contextual requirement accords with How People Learn II, which emphasizes interactions among prior knowledge, learning strategies, materials, goals, and cultural contexts (National Academies of Sciences, Engineering, and Medicine, 2018). A reduction in curriculum size alone is consequently not evidence of improved cognitive development. The claim must concern what learners become able to do and the conditions under which they can do it.

An operational test can begin by defining a target task family, an eligible learner population, an instructional budget, and an adequacy threshold before comparing material sets. For illustration, adequacy might require at least 80% of learners to reach a specified reasoning standard on parallel unfamiliar tasks. The percentage is a design example, not an established educational benchmark; its justification and the acceptable loss margin must be set for the actual context.

Planned removal comparisons must target relations as well as materials. Before comparison, construct a relation-to-carrier matrix: each target relation is linked to every poem, diagram, example, teacher prompt, and practice activity that could teach it. Removing one poem while another activity teaches the same relation tests that poem's contribution, not the relation's necessity. To test omission of a relation, remove or redesign its planned instructional routes while preserving other target content and documenting residual exposure. If that separation is infeasible, necessity remains untested. Learners' existing knowledge cannot be deleted; the estimand concerns omission from the specified instruction.

The time-filling rule is part of the comparison. If the removed material is replaced with a neutral activity that does not teach the target relation, the contrast examines omission within that stated activity budget. If the time is used for another relevant carrier, explanation, or practice opportunity, the contrast examines relative instructional value against that substitute. Neither route creates a pure absence-of-input condition; time, instructor attention, workload, and additional exposure should be documented rather than assumed equivalent. Omission and replacement comparisons should be reported separately.

An inclusion-minimal set is adequate while none of its tested proper subsets remains adequate under the same criteria. For a relation-level test, compare full instruction with a mapped omission condition; retain a separate carrier-replacement comparison when relevant. Check compensatory cues, overlapping relations, workload, and teacher explanations. A

meaningful loss after omission supports instructional importance within that tested pathway, not logical or universal necessity. One-at-a-time removals are insufficient when relations are redundant or interact; selected joint removals and replication would be needed. A precise loss within an acceptable margin supports local dispensability; imprecision leaves the question unresolved.

4.3 A micro-worldview and the position of the self

A micro-worldview is an unmeasured pedagogical heuristic for connecting aspects of a learner's minimal cognitive core across self, other people, environment, action, and uncertainty. The core names the learner's developing organization of usable relations; micro-worldview names a proposed way of discussing connections among them, not an additional psychological faculty or a scored outcome. Table 1 illustrates possible relations. The seven activities in Section 6 describe ways understanding may be exercised; they are neither components of a validated micro-worldview scale nor a developmental sequence.

Table 1. Candidate relations for an initial micro-worldview

Relation	Illustrative question
Observation and inference	What did I actually observe, and what did I add to it?
Action and consequence	What might change if the action or circumstances change?
Self and other perspectives	What could another person notice or value differently?
Evidence and uncertainty	What would help distinguish competing explanations?
Understanding and revision	What new information would require me to reconsider?

Beginning with the self means recognizing the learner as a situated observer and agent who experiences consequences. It does not grant personal preference authority over evidence or make other people peripheral. The educational task includes learning that one's perspective is partial and that others possess experiences, interests, and capacities to disagree. Personal agency and perspective-taking must therefore develop together.

The metaphor of a world model can help discuss connections between representations and experience, but an analogy with AI does not establish mechanistic equivalence. The present proposal makes no inference about human consciousness from AI architecture. It also leaves open whether one coherent micro-worldview is always desirable; productive understanding may require several context-sensitive models whose tensions are acknowledged rather than prematurely resolved.

4.4 Relationship to existing constructs and discriminating tests

Spiral curriculum design revisits a subject's underlying structures at increasing complexity (Bruner, 1960). Cognitive Seeds shares revisitation and progressive elaboration; its additional specification is to map candidate relations to all instructional routes and test local omission and generative use. This is an operational emphasis, not a property absent from spiral curricula.

Threshold concepts concern understandings that transform participation in disciplinary thinking and can be troublesome to acquire (Meyer & Land, 2003). A seed need not be a disciplinary gateway or a transformative threshold. Its selection is provisional and its instructional value can be tested without claiming a discontinuous transformation. Threshold-concept instruction can also promote transfer; transfer alone cannot distinguish the frameworks.

Big-ideas and backward-design approaches already connect central understandings, transfer, assessment, and instructional activities (Wiggins & McTighe, 2005). Cognitive Seeds does not claim to invent relation-centered curriculum or alignment. Its proposed increment is a compact, explicitly mapped starting set with context-bound dispensability tests and a separate test of learning from new resources. Whether those requirements improve a design must be evaluated.

Conceptual-change research examines how prior conceptions are restructured when learners encounter reasons to reconsider them (Posner et al., 1982). Cognitive Seeds shares attention to prior understanding and warranted revision, but does not offer an alternative account of the mechanism of conceptual change. It supplies selection and assessment commitments that can be implemented within such an account.

Cognitive Seeds is an integrative design framework, not an explanatory theory or a psychological kind distinct from schemas, mental models, or conceptual understanding. Its vocabulary specifies what designers should select, make observable, and compare. A substantive contribution would consist in better specified decisions, useful predictions about a stated design contrast, or improved outcomes beyond an adequate comparison. Renaming familiar teaching is not such a contribution.

Table 2 translates the commitments into comparisons. Recall-oriented and unconnected-content conditions can establish a contrast but are weak tests of added value. A stronger active comparison should also provide explanation, feedback, and transfer opportunities. If the

framework adds no practical guidance or meaningful benefit beyond such instruction, it should remain an organizing vocabulary rather than be promoted as an independent theory.

Table 2. Discriminating comparisons for the proposed design commitments

Design contrast	Expected observation	Interpretive limit
Recall-matched teaching with or without explicit boundary cases	The boundary-case condition better distinguishes when a learned relation applies and when it does not.	Supports boundary-case teaching; does not uniquely establish a seed mechanism.
Time-matched connected relation maps versus the same facts taught separately	Connected maps support more justified links and new questions on unfamiliar cases.	Tests relational organization beyond fact coverage; schema approaches may predict the same result.
Alternative carriers mapped to the same target relation	Learners can explain and appropriately revise the relation when the familiar carrier is replaced.	Examines carrier dependence; transfer across cultures still requires separate evidence.
The full design versus an active elaboration and transfer comparison	Prespecified outcomes improve beyond a well-supported active baseline at comparable cost and time.	An advantage supports incremental design value; identifying a distinct mechanism requires further analysis.

These comparisons also create failure conditions. Recall without transfer weakens the claim that a carrier has functioned as a seed; indiscriminate application weakens the claim of revisability; and no incremental benefit over a strong active comparison weakens the claim that the framework adds instructional value. None of these outcomes warrants denying that the learner has other forms of understanding.

A testable incremental prediction is that an explicit relation map plus planned replacement of familiar carriers will improve justified application to unfamiliar carriers and recognition of boundary cases beyond a specified active comparison. Match target relations, instructional time, examples, explanation opportunities, and feedback, and document the actual differences. Schema-based, big-ideas, and conceptual-change instruction may predict the same outcome. If those features already occur in the comparison, the proposed contrast is absent. A precise lack of advantage weakens the added-value claim; an imprecise estimate leaves it unresolved. Success would support the tested design package, not establish a unique seed mechanism.

5. Cultivation through progressive elaboration

The practical proposal is to revisit a limited selection of carriers while changing what the learner is asked to do. Early encounters can establish vocabulary and familiarity. Later encounters can invite explanation, comparison, alternative cases, reflection, and revision. Familiarity becomes a resource for inquiry rather than an endpoint.

Chi et al. (1989) associate self-generated explanations with understanding and reduced reliance on examples in a mechanics-learning context. Chi and Wylie (2014) distinguish modes of engagement through the ICAP framework, including constructive and interactive activity. These studies provide reasons to investigate learner-generated elaboration, but they do not validate Cognitive Seeds, poetry as its carrier, or a particular developmental schedule. The proposal must show what it adds to existing approaches.

Its candidate contribution is a design constraint: selected starting materials should have an explicit map of relations, opportunities for learner-generated elaboration, and planned occasions for revision and transfer. Without such a map, the intervention could simply rename familiar reflective teaching. Comparison studies must therefore specify whether the proposed benefit comes from content selection, repetition, explanation prompts, instructional time, or their combination.

The author describes a source curriculum using 26 Chinese quatrains across themes including emotions, values, nature, society, and science (Gao, 2024). Its six options—recitation, understanding, learning, appreciation, exploration, and thinking—illustrate progressive elaboration, not verified age-bound stages. Poetry is one candidate carrier; figurative meaning, cultural familiarity, translation, and interpretation accuracy can affect its suitability. Stories, pictures, games, and practical activities are alternatives, not demonstrated equivalents.

The six instructional levels organize teaching choices, while the seven activity dimensions in Section 6 describe candidate responses. They do not require a one-to-one mapping: an appreciation activity may involve connection, judgment, and reflection, and a relation diagram may involve observation, reasoning, and revision. These organizational schemes should not be interpreted as validated developmental stages.

In technical or non-poetic settings, a short program, a mathematical relation, a diagram, or a paradox could be considered as a carrier when it meets the same accessibility, elaboration, and revision criteria. For example, changing a program's input can test the limits of a rule. This is a proposal for alternative materials, not evidence that the framework generalizes across media or disciplines.

The Three Little Pigs illustrates how cognitive and values-education objectives must be separated. A cognitive task might ask learners to compare explanations of a building's stability, vary resources, or identify evidence that could distinguish alternatives. A values discussion might consider responsibility, fairness, or cooperation. These objectives need separate instructional statements and evidence. A reasoned moral disagreement with the teacher must not lower a reasoning score; agreement with a preferred moral is not evidence of understanding a causal relation.

Progression also requires adaptation. A novice unable to explain an event may need demonstration or vocabulary support before being asked to generate alternatives. Extra friction can exhaust rather than assist. The framework therefore treats challenge as something to calibrate against task demands and prior understanding, not as an educational good in itself.

6. From proposed capacities to observable activity

Seven verbs organize candidate observations: observe, connect, question, reason, judge, reflect, and revise. Table 3 offers task-specific operational candidates. They describe activities, not seven established faculties or the measured structure of a micro-worldview. Each requires an elicitation opportunity and task-specific checks. A behavioral indicator concerns performance under stated conditions, not direct access to an internal cognitive state.

Table 3. Candidate activity dimensions and observable indicators

Dimension	Working definition	Candidate indicator
Observe	Attend to relevant features and distinguish observation from inference.	Separates stated evidence from an added assumption.
Connect	Relate information through a meaningful structure.	Explains a relevant similarity and its limits in a new case.
Question	Identify a specific gap or uncertainty worth investigating.	Asks for information that could distinguish explanations.
Reason	Develop an inference from stated grounds.	Links premises to a conclusion and considers a counterexample.
Judge	Assess a claim using relevant criteria and evidence.	Identifies a specific unsupported inference without rejecting sound claims.
Reflect	Examine one's own interpretation, strategy, or confidence.	Names an assumption and explains how it influenced the answer.
Revise	Change an interpretation in response to relevant grounds.	Explains what changed, why, and what remains uncertain.

Metacognition supplies an important connection to the last two dimensions. Flavell (1979) distinguishes knowledge about cognition from experiences of monitoring it. A learner's statement of confidence, an account of a strategy, and an actual correction should

accordingly be recorded separately. Reporting reflection is not necessarily accurate monitoring, and changing an answer is not necessarily a justified revision.

The dimensions can overlap. Identifying a flawed causal inference may involve observation, questioning, reasoning, and judgment. Until their measurement properties are established, summing them into one score could conceal such overlap and create unjustified precision. A preferable initial report presents task scores and qualitative evidence by dimension, explicitly marking activities that a task did not elicit.

No empirical study presented here establishes discriminant validity for these seven dimensions. The first studies should report dimension-specific evidence and scores without a seven-dimension total. Code a dimension as NE, not elicited, when the task provides no meaningful opportunity to demonstrate it. For example, a one-shot evaluation with no initial response or subsequent new evidence cannot adequately elicit revision and should receive NE for that dimension, not zero. A separate planned step supplying conflicting evidence can make revision assessable.

When the task does elicit the activity and a valid response shows no relevant demonstration, a rubric score of zero may be appropriate. Missing responses, technical failures, and not-elicited activities require distinct codes and prespecified handling. When the same passage supports more than one dimension, coders should retain the shared evidence identifier and avoid treating the resulting codes as independent observations. A narrower task-specific endpoint can combine explicit reasoning criteria; that does not validate a global cognitive-capacity score.

A preliminary rubric can distinguish no relevant demonstration, a partially justified demonstration, and a coherent demonstration supported by task-specific evidence. It needs examples of acceptable alternatives and common false positives. Assessors should score anonymized responses without knowing the experimental condition where feasible, independently double-score a planned subset, and document disagreement. Agreement between assessors would establish consistency of scoring, not the validity of a theory of cognition.

Nonverbal evidence also needs task-specific anchors. For a tower-stability task, learners could choose between two unfamiliar base configurations, construct a stable alternative after a failure, and point to or demonstrate the changed relation. A candidate rubric should specify which choices and changes count as relevant, which can be explained by copying or motor skill alone, what assistance is permitted, and when the task did not elicit the intended activity. A successful construction is not automatically evidence of causal understanding. Scores from different response modes should not be assumed interchangeable without checking their demands and interpretation.

7. Implications for interaction with generative AI

The framework proposes that a learner should sometimes formulate an initial interpretation before consulting AI, then compare the response with that interpretation and test the resulting understanding in another setting. This sequence creates opportunities to make assumptions and revisions visible. It does not establish that prior formulation is

necessary for every useful AI interaction, or that an unrestricted interface always prevents learning.

The prior-formulation, AI-comparison, and independent-application sequence remains a design hypothesis. Self-explanation research (Chi et al., 1989) motivates eliciting a learner's own explanation. Bastani et al. (2025) motivates distinguishing a general-purpose assistant from a teacher-informed tutor and measuring unassisted outcomes. Neither study directly tests the sequence proposed here. An immediate-access comparison must therefore use a specified AI role, not be equated with every form of unrestricted generative AI use.

Kapur (2016) distinguishes productive and unproductive combinations of initial performance and later learning. The relevant rationale is that an initial generation attempt can prepare learners to benefit from subsequent instruction, even when the attempt is unsuccessful; subsequent instructional consolidation matters. This does not establish that any delay, difficulty, or failed attempt is useful. Prior formulation here need not involve failure, and a live AI response may not supply effective consolidation. Thus productive failure motivates a hypothesis and checks on instructional support, workload, and prior knowledge; it does not validate this AI sequence.

Three outcomes deserve separate attention: the quality of the assisted artifact, the quality of the learner's contributions during interaction, and subsequent performance on a comparable task without the same assistance. A fluent final response can coexist with weak independent explanation; an imperfect assisted response can coexist with useful questioning and correction. These combinations make a single productivity metric inadequate for the educational question.

Independent tasks must still be defined relative to the target capability. A no-AI assessment can examine whether participants retain an argument structure or detect an unsupported inference, while allowing ordinary accessibility supports that do not supply the tested reasoning. An AI-free task cannot establish total cognitive autonomy. Conversely, performance with a tool may be valuable even if the learner is not expected to reproduce every operation unaided.

For adults, the emphasis may be on exercising and revising existing understanding rather than constructing an initial educational foundation. For children, content, support, language, and consent procedures require separate consideration. Evidence from an adult analytical pilot cannot validate a preschool curriculum or establish developmental benefits. The relationship between these applications is a research question, not an assumed generalization.

8. Research propositions and a staged evaluation agenda

8.1 Four propositions

Proposition 1 concerns relational elaboration. Under comparable instruction, explanation, comparison, and revision opportunities are expected to support better independent application than carrier reproduction alone. A stronger test compares the fully specified package with existing active instruction. No practically meaningful advantage with adequate precision would weaken its incremental instructional-value claim in that setting.

Proposition 2 concerns compactness. A compact, connected starting set may support integration when it supplies adequate domain knowledge. Removing material that learners need should weaken rather than improve application. This proposition rejects an unrestricted claim that less knowledge is better and requires comparison with a well-designed broader curriculum.

Proposition 3 concerns interaction structure. Prior formulation and explicit comparison with an AI response may improve specified subsequent reasoning activities relative to direct AI use. The effect could be absent, negative, or limited to learners with sufficient prior knowledge. A time-matched preparation condition is needed to distinguish additional thinking time from the particular structure of the intervention.

Proposition 4 concerns generative use. After initial instruction, learners should be able to ask a diagnostic question, use new information to learn an extension, and revise only when the evidence warrants it. Independent transfer without improved learning from new resources would support a narrower application claim, not the fuller generative claim. Rigid repetition, indiscriminate revision, or gains confined to a familiar response template would count against the proposed interpretation. A lack of incremental advantage on PFL (Preparation for Future Learning) tasks—even when accuracy is already sufficiently high—serves to weaken the claim for generative use.

8.2 Feasibility before effectiveness

An internal walkthrough with two or three willing reviewers checks readability, timing, elicitation opportunities, AI fidelity, and scoring disagreements. It cannot estimate effectiveness or stable reliability. Reviewers familiar with intended flaws may test usability, but their detection scores are not unbiased outcome evidence. Repeating a case across conditions likewise introduces exposure effects. Keep calibration responses separate from later outcome data and record changes before freezing materials, prompts, anchors, and planned contrasts.

Curriculum evaluation and AI interaction evaluation should initially be separated. A curriculum study can compare teaching methods with the same carriers, content, and instructional time. A later comparison can examine alternative carriers or curriculum breadth. An interaction study can hold the learning material constant while comparing direct AI use, time-matched unstructured preparation followed by AI, and structured prior formulation followed by AI. An eventual adequately powered factorial design could examine how curriculum and interface features interact.

8.2.1 Interaction conditions and the tutor's role

An illustrative adult comparison uses common three-minute reading followed by a ten-minute interaction window. Immediate access provides ten minutes with the specified AI tutor. Free preparation provides four minutes of unrestricted private notes without AI, then six minutes with the tutor. Structured preparation provides four minutes to state an initial explanation, distinguish evidence and assumptions, identify an uncertainty, and name evidence that could change the explanation, then six minutes with the same tutor. All arms receive the same final-response instruction; the two preparation arms also receive the same

comparison instruction after AI access. Timings remain calibration choices, not an adopted protocol.

The primary structure contrast is structured versus free preparation, matching preparation time, AI exposure, final instructions, and materials. Record spontaneous structuring in free notes rather than prohibit it. Immediate access estimates an overall package contrast that also changes AI exposure; it cannot isolate friction. Log actual active time, tutoring turns, reading time, interruptions, and deviations. Equal scheduled time does not ensure equal cognitive effort. A short workload rating is supplementary and cannot by itself identify a mechanism.

For this comparison, the AI is a reasoning tutor rather than an answer generator. A candidate common instruction is: “Help the learner examine the report using questions and requests for justification. Distinguish reported facts from missing information. Do not invent allocation procedures or other facts. Do not provide a ready-made critique or reveal the target flaw. Ask the learner to consider plausible alternatives and the limits of the conclusion; acknowledge uncertainty.” This instruction requires auditing and testing; it does not guarantee compliance. Use the same tutor instructions and capabilities in all arms, disable browsing and external tools in the bounded task, and preserve complete task-relevant conversations.

A separate operator manual must fix the allowed user prompts, turn or time limits, resources, handling of clarification requests, and stopping rules. Any named interface cannot be assumed to be implemented by these general principles; any named procedure needs its own accepted specification. Internal walkthrough participants should inspect whether tutor behavior changes across the preparation conditions before an outcome pilot begins.

These timings specify a controlled interaction component, not a complete lesson. Any implementation in a longer class must document surrounding instruction and keep it comparable across conditions. No lesson timetable or related essay substitutes for a frozen experimental protocol.

Use short adult analytical reports with enough information to distinguish observations from causal and adoption claims. Include both an overreaching report and one with a justified, appropriately limited conclusion. Lack of reported randomization is missing information, not proof of selection bias. An immediate test can be informative without establishing durable learning. Pretest readability and difficulty with people outside the outcome sample. Use parallel baseline items without disclosing the study case’s target inference, and record prior domain knowledge, reading proficiency, and AI experience.

8.2.2 Independent tasks and follow-up

The first delayed endpoint is a new task approximately 24 hours later without generative AI. Change surface content and response format while retaining the reasoning demand. Section 8.3 describes the non-isomorphic format; Section 8.4 adds preparation for future learning. These are different endpoints: transfer asks what can be done on a new problem, whereas preparation for future learning asks how participants use a subsequent learning opportunity.

A follow-up around one week later can add durability information, but no interval alone establishes retained understanding. Fix allowable timing windows, reminders, accessibility supports, and resource restrictions before collection. Record deviations and possible intervening practice; remote compliance cannot simply be assumed. Report attrition by condition and prespecify missing-data analyses. Neither a 24-hour nor a one-week difference establishes general cognitive preservation, augmentation, or substitution.

8.3 Measurement and interpretive boundaries

For the proposed interaction study, the primary endpoint is the prespecified, task-specific 24-hour non-isomorphic transfer score; the primary contrast is structured versus free preparation. A preparation-for-future-learning score is an additional endpoint, not a post hoc replacement for an unsuccessful primary test. Immediate output quality, false criticism, confidence, workload, protocol fidelity, and later follow-up are secondary or process measures. Preregister the endpoint hierarchy and multiplicity plan before an outcome study; the present article provides no validated scoring instrument.

8.3.1 Assessment alignment and format familiarity

A candidate transfer task presents a new domain with two causal diagrams, alternative explanations, and possible follow-up comparisons. Participants choose which claim the report supports, select a diagnostic comparison, and predict a result under a changed condition, giving brief reasons. Instructions do not repeat the preparation categories or ask for the same four-part template. A task-specific 0–6 score can assign 0–2 for each of three targets: correctly limiting the inference, giving a case-grounded explanation or prediction, and selecting evidence that distinguishes alternatives. Each target requires its own correct, partial, and incorrect anchors. These targets share the intended reasoning content with instruction; changing format does not eliminate all alignment.

Before collection, all groups should receive equal practice with the assessment's diagram and selection format using unrelated content and no instruction in the target inference. This reduces format unfamiliarity without teaching the test solution. Use blind review to check that neither wording nor scoring rewards reproduction of the structured prompt. Freeze task-specific alternatives, sound-claim anchors, and opportunity checks after calibration. A bounded 0–6 score is not a seven-activity total; similarly structured scores across topics are not automatically interchangeable. Prespecify double scoring and adjudication.

8.3.2 Exposure, justified revision, and study controls

Separate report-flaw detection from AI-output evaluation. For detection rates, define diagnostic flaws and sound claims in advance, use justified identifications as the numerator, and state the opportunity denominator. A common audited AI response containing a controlled error, paired with a sound response, can test evaluation under equal exposure; it does not estimate a model's natural error rate. Keep this controlled stimulus separate from live tutor conversations. Expert review of live outputs should record natural errors and direct disclosure of the target flaw, including wording, turn, timestamp, and whether disclosure preceded the participant's identification.

To distinguish justified revision from copying or deference, retain a time-stamped initial response, the AI text actually seen, and a final response with reasons for changes. In a separate common-response check, present evidence that warrants a change on one item and an incorrect AI recommendation on another; score both evidence-sensitive revision and justified resistance. Add a new application item requiring a prediction or counterexample that the AI has not supplied. Verbatim or near-verbatim uptake is an exposure flag, not proof of absent understanding; changing one's answer or paraphrasing the model is likewise insufficient. Blind raters should assess the grounds and the new application separately from surface similarity.

Direct tutor disclosure is a protocol-fidelity event, not an automatic participant exclusion. Prespecify a primary analysis retaining randomized participants in their assigned conditions and descriptive disclosure rates, with clearly labeled sensitivity analyses if justified. Because disclosure is a post-assignment event, conditioning on it can bias a causal comparison. Log it rather than silently relabel the trial as successful mediation.

An exploratory pilot should prioritize completion, usability, fidelity, score distributions, and uncertainty, with its sample justified by those goals. Confirmatory effectiveness requires a separate justified sample-size calculation, a locked primary contrast and endpoint, random assignment or another defensible identification strategy, and prespecified exclusions and missing-data handling. Use baseline parallel-task performance and domain knowledge for balance checks and prespecified adjustment; do not select covariates after seeing favorable results. Calibrate task difficulty, avoid ceiling and floor effects, and keep calibration participants out of the confirmatory sample.

A delayed score difference is an observation. Calling it retention, augmentation, or substitution requires additional support. Prompt alignment, prior knowledge, AI disclosures, exposure, motivation, and task difficulty remain alternative explanations. If structured preparation does not outperform time-matched free preparation with adequate precision, the structure-specific claim is weakened. If gains disappear when format changes or merely reproduce AI wording without successful application, the broader interpretation is weakened. Imprecise null estimates establish neither absence of benefit nor equivalence.

Permit favorable, unfavorable, mixed, and inconclusive results. Report uncertainty and process deviations alongside outcomes. Freeze model identity or service version, date, system and user instructions, settings where exposed, and tools enabled. Use separate sessions without cross-participant memory and record any provider update. If the model cannot be pinned, hold collection windows narrow, log changes, and prespecify sensitivity analyses or a pause rule. Do not pool a substantive model change silently. A model-mediated result applies to the tested instructional arrangement; it is not a universal property of AI.

8.4 Preparation for future learning and weakening results

Bransford and Schwartz (1999) argue that transfer assessment should include how prior learning prepares people to learn from later resources, not only what they can reproduce unaided. Here, preparation for future learning is an additional endpoint: using unfamiliar information to extend or revise a relation after initial instruction. It is conceptually distinct from the primary delayed-transfer score. It does not measure general readiness to learn.

A proposed task first presents a new domain and a short baseline prediction, then offers a fixed menu of information sources under an equal selection budget. Sources include diagnostic comparisons, useful background, and plausible but nondiagnostic details; their accuracy and relevance are independently checked. Record which information the learner requests and why. All learners subsequently receive the same concise new lesson, including a boundary condition not supplied in the original teaching. They then solve a different case without AI, predict an outcome, explain how the new information matters, and state when their initial relation would need modification. Alternate forms and counterbalancing are needed to limit item and order effects.

Candidate secondary scores are relevance of the information choice, integration of the new lesson in a new prediction, and warranted revision at the boundary case, each with task-specific anchors. Record before-lesson performance, but do not treat an unadjusted gain score as a causal mechanism. A future outcome study should prespecify between-condition contrasts, baseline adjustment, acceptable alternative answers, and uncertainty reporting. Pilot item difficulty and score interpretation separately; a composite should not be assumed valid merely because its parts are coded 0–2.

The fuller generative claim would be weakened by accurate delayed reproduction accompanied by poor resource use, no meaningful advantage in learning the extension, failure to recognize its boundary, or an advantage confined to the original carrier or template. With adequate precision, no added benefit over active instruction would weaken the framework's incremental-value claim. Imprecise results remain inconclusive. Even favorable preparation-for-future-learning results would support the tested instructional design and target relations, not uniquely identify a cognitive seed mechanism.

9. Limitations and ethical considerations

The framework integrates existing ideas and may add no explanatory power beyond them. The empirical issue is whether its explicit relation maps, omission rules, and evaluation commitments improve a specified design or outcome. Its micro-worldview remains an unmeasured organizing heuristic; task-level findings cannot validate that broad network. Favorable results would support tested components or a design package, not demonstrate a distinct mental object named a cognitive seed.

A second limitation concerns content selection. A supposedly minimal foundation can reproduce the selector's cultural assumptions or moral preferences. Materials should expose alternative interpretations and disclose normative choices. Learners' disagreement should not be coded as cognitive failure. Multiple localized implementations may be preferable to one centrally defined canon.

Cross-language implementation must revalidate the mapping from carrier to target relation, including implied values, figurative meanings, familiarity, and possible alternative interpretations. Translating a Chinese quatrain does not establish that it elicits the same relation or level of difficulty elsewhere. Local co-design, interpretation checks, and adapted examples should precede comparisons. Claims of cross-cultural generality would require evidence that the tasks and scoring criteria remain interpretable across the populations studied.

Third, a demand for explicit explanation can favor verbal fluency and overlook embodied, practical, or nonverbal understanding. Assessment should allow appropriate demonstrations, diagrams, or oral responses and investigate language and accessibility effects. A learner's inability to produce the expected wording cannot by itself establish lack of understanding.

Any participant study should obtain appropriate institutional ethics review or a documented determination before recruitment, secure informed consent, and minimize identifiable data. Studies involving children require developmentally appropriate assent and guardian permission where applicable. Planted errors should appear only in approved, low-stakes materials and be explained afterward. Interaction logs should record task-relevant actions without claiming to reveal private mental states. Participation in discussion does not imply agreement to recruitment, recording, or an evaluator role.

A child-specific feasibility starting point could use a small group, one brief session, and no AI, examining whether instructions and materials are understandable and whether the activities can be completed. Record assistance, discomfort, and reasons for stopping. Such a walkthrough does not estimate educational effects or establish developmental stages, and it should precede considering an AI interaction study with that age group.

10. Conclusion

Cognitive Seeds is an integrative design framework for specifying relations, instructional carriers, and evidence of learner use. Its distinctions prevent recall from being equated with understanding, a supplied knowledge set with a learner's cognitive core, or an assisted product with independent capacity. Micro-worldview remains an unmeasured pedagogical heuristic rather than an additional outcome.

The next step is a bounded feasibility check of the relation map, materials, tutor instructions, non-isomorphic tasks, and scoring anchors. Later studies can test independent application and learning from new information against credible existing instruction. No new mechanism or educational benefit is established here. The framework should be retained, simplified, or rejected according to whether its explicit commitments add useful guidance and withstand those comparisons.

Declarations

Data availability: This conceptual paper reports no newly collected participant data. The examples and proposed protocols are illustrative.

Ethics: No human-subject study is reported. The research agenda describes procedures for possible future studies, not an approved or completed experiment.

Funding: No funding is reported.

Author's Position and Interests: This framework was proposed by John Gao, whose works on foundational education are authored under the pseudonym "Si Jiu". The author maintains an academic/intellectual interest in the development of this proposal.

AI Assistance: OpenAI's ChatGPT model was utilized during this revision process to assist with drafting the English text, refining concepts, conducting literature searches,

verifying references, and preparing the document. No participant observation was conducted, and no new data or research findings were generated. This statement does not imply the introduction or evaluation of any specific "study tutor" mechanism.

Literature and Sources: The literature search and synthesis process was selective rather than systematic. The selected literature is based on conceptual relevance, as outlined in Section 2. The revised interpretations of the literature, as well as any newly added references, have been verified against records published by the respective publishers, relevant institutions, or the author himself. John Gao bears responsibility for the assumptions upon which this framework is built, as well as for its final arguments.

References

1. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122
2. Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. Review of Research in Education, 24, 61–100. https://doi.org/10.3102/0091732X024001061
3. Bruner, J. S. (1960). The process of education. Harvard University Press.
4. Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1
5. Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4), 219–243. https://doi.org/10.1080/00461520.2014.965823
6. Clark, A., & Chalmers, D. J. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7
7. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. American Psychologist, 34(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906
8. Gao, J. (2024). New Education: Cultivating Seeds of Cognition (English only): ChatGPT Era (CS1). Amazon. https://www.amazon.com/gp/product/BODM6HJT8N
9. Gao, J. (2025). A Guide for Planting the Future of Education: Cognitive seeds textbook toolkit (CS1 4). Amazon. https://www.amazon.com/gp/aw/d/B0FBKGMCHM
10. Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educational Psychologist, 51(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

11. Meyer, J. H. F., & Land, R. (2003). Threshold concepts and troublesome knowledge: Linkages to ways of thinking and practising within the disciplines (Occasional Report 4). Enhancing Teaching–Learning Environments in Undergraduate Courses Project, University of Edinburgh. https://pressbooks.atlanticoer-relatlantique.ca/app/uploads/sites/803/2021/07/ETLreport4.pdf
12. National Academies of Sciences, Engineering, and Medicine. (2018). How people learn II: Learners, contexts, and cultures. The National Academies Press. https://doi.org/10.17226/24783
13. Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education, 66(2), 211–227. https://doi.org/10.1002/sce.3730660207
14. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002
15. Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. Perspectives on Psychological Science, 10(2), 176–199. https://doi.org/10.1177/1745691615569000
16. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
17. Wiggins, G., & McTighe, J. (2005). Understanding by design (Expanded 2nd ed.). Association for Supervision and Curriculum Development.