Vibe Compiler: A Research-Logic Synthesis Tool That Runs without Prompt Engineering —Toward Enhancing Metacognition for Sustaining Agency in the Age of Generative AI—

Riichiro Mizoguchi[a,*], Tomoki Aburatani[b], Kento Koike[c], Machi Shimmei[d]

[a] *Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan*
[b] *Osaka Metropolitan University, Osaka, Japan*
[c] *Kanagawa University, Yokohama, Kanagawa, Japan*
[d] *Tohoku University, Sendai, Miyagi, Japan*

arXiv:2608.05545v1 [cs.CY] 6 Aug 2026

Abstract

Generative AI used as a capable servant has greatly accelerated intellectual work, but it also risks eroding human epistemic agency by encouraging uncritical acceptance of AI-generated reasoning. This creates a need for mechanisms that preserve human agency by augmenting metacognition during AI-assisted intellectual work. To address this, we propose the Synthesis–Analysis Reciprocity Model, which views intellectual construction as a reciprocal interaction between two cognitive functions: Synthesis, which selects and combines components in the artifact, and Analysis, which critically evaluates them against objective indicators and constrains the subsequent synthesis.

Grounded in this model, we present a Vibe Compiler, a research-logic compiler that helps researchers transform vague ideas (Vibes) into coherent research logic. The system attempts to compile these ideas using a research paper ontology consisting of sixteen academic parameters. Compilation failures are interpreted as indicators of missing logical components. Instead of filling these gaps autonomously, the system prompts researchers with reflective questions, encouraging them to develop the missing reasoning themselves.

The framework further characterizes the origins of structural gaps along two orthogonal dimensions—cognitive function (Synthesis vs. Analysis) and executing agent (human vs. AI)—yielding four types of origin and providing a principled way to identify where breakdowns in intellectual construction originate. Above all, the design adopts the type in which the AI probes its own synthesized output and thereby excites the human's metacognition. This design encourages researchers to remain "managers" of AI-assisted reasoning who critically direct and validate AI outputs rather than becoming passive "makers" of the output.

Experience with the prototype built on NotebookLM and Gemini shows that effective

*Corresponding author

Email addresses: mizo@jaist.ac.jp (Riichiro Mizoguchi), aburatani.tomoki@omu.ac.jp (Tomoki Aburatani), kento@koike.app (Kento Koike), machi.shimmei.e6@tohoku.ac.jp (Machi Shimmei)

AI-assisted reasoning depends less on sophisticated prompting than on the structure of the knowledge provided to the AI. The proposed framework operates at two complementary levels, cultivating learners' metacognitive reflection and inspecting researchers' construction of logic, and we position it as a self-applying epistemic operating system that uses generative AI to mitigate the crisis of human agency introduced by AI itself. This paper was itself developed using the proposed Vibe Compiler.

Keywords: generative AI, epistemic agency, metacognition, evaluative judgement, synthesis–analysis reciprocity, paper ontology

---

1. Introduction

1.1. Background: The "Crisis of Agency" in the Age of Generative AI

The domain this study addresses is the construction and inheritance of knowledge within the scholarly information infrastructure. Scholarship, properly understood, consists in scholarly accumulation, in reading the context of prior work and building new logic upon it through critical examination, and it is this discipline of accumulation that has turned individual results from transient reports into contributions to a body of knowledge (Scardamalia & Bereiter, 2014). The rapid spread of generative AI (hereafter GenAI) is nevertheless reducing this very process of construction to the efficient processing of information. The higher-level problem is thus located in how human-driven knowledge creation, that is, epistemic agency (Bandura, 2006), is to be secured under digital transformation. Epistemic agency denotes the capacity to decide for oneself what to ask, what to take as grounds, and what to judge valid, rather than to receive knowledge passively.

That this crisis is more than an abstract concern has been confirmed at several levels. Where the trial and error and the reflection that learning requires, namely productive struggle (Hiebert & Grouws, 2007; Warshauer, 2015), are displaced by the automatic generation of AI, the very path toward deep understanding is severed. Users who place greater trust in GenAI indeed expend less cognitive effort on critical thinking (Gerlich, 2025; Lee et al., 2025), and delegating the act of writing is accompanied by a decline in neurophysiological engagement (Kosmyna et al., 2025). Once such delegation becomes routine, the outsourcing of thinking and writing settles into cognitive offloading (Risko & Gilbert, 2016), and even the attribution of responsibility for errors turns opaque (Dwivedi et al., 2023). A user who swallows the output whole while responsibility remains unlocated loses the occasion to exercise the evaluative judgement (Bearman et al., 2024; Tai et al., 2018) through which the quality of information is governed, and falls into the paradox of producing more fragile artifacts under AI assistance while growing more confident in what they have produced (Perry et al., 2023). Because factual error (Huang et al., 2025; Ji et al., 2023) and excessive agreement with the user (Sharma et al., 2023) are unavoidable in principle, individual carefulness cannot dissolve these effects. The problem resides not at the level of the user's frame of mind but at the level of the design philosophy of the support system.

Since entanglement with AI is irreversible, what must be asked is not whether AI should be used but how the human role is to be redefined. Cox (2024) places the purpose

of education in the age of AI in a transformation of the learner's role, calling not for learners to remain makers of knowledge but for their elevation into managers who critically control and audit the information AI generates, and further into inforgs (Floridi, 2014) who exercise the initiative as an organic part of the AI environment. What is decisive here is that agency as a manager is sustained only by continually honing metacognition (Flavell, 1979). Agency and metacognition stand in a mutually complementary relation, in that by continually honing metacognition one retains agency as a manager of information without surrendering the initiative to AI. The core of the capability that should remain on the human side in the age of AI is therefore metacognition itself, and metacognition is likewise what a support system must protect and drive above all else.

1.2. The Problem to Be Solved

Current use of GenAI tends to end in a catalogue of results of the form "we built a system and it worked," while the accumulation of logic that contributes to scholarly progress is neglected. GenAI converts a vague intuition, a "Vibe," into a working artifact at once. "Vibe coding" (Karpathy, 2025; Sarkar & Drosos, 2025), in which one forgets that the code exists at all and gives oneself over to the mood, symbolizes how readily the production of an artifact separates from the construction of the logic that makes it valid. When that separation occurs, the user regresses into a pseudo-maker who leaves the construction of logic to AI and serves only as its signatory.

This regression appears isomorphically in two layers, that of the learner and that of the researcher. In arithmetic problem posing, a learner reports that "I made a problem that asks for the amount to pay when you buy an apple for 100 yen and a 10% consumption tax applies, and since it needs a calculation it is fairly difficult." The number of operation steps is nevertheless one and the number of unknowns is likewise one, so the subjective assessment of difficulty diverges widely from the actual structure. The learner, however, holds no means of noticing this discrepancy. The situation is no different on the side of the researcher, who reports only that a prototype "worked" and proceeds toward writing a paper while the significance, the intended beneficiaries, and the limitations of existing methods remain undescribed (Null). In both layers a subjective sense of accomplishment conceals an objective deficiency of structure. This is isomorphic to the known phenomenon in which perceived productivity diverges from actual understanding under AI assistance (Barke et al., 2023; Vaithilingam et al., 2022).

The problem this study sets out to solve is therefore how to design and realize a support mechanism that hones metacognition while preserving the human's authority to decide, in the process of converting vague intuition into scholarly logical rigour. What the problem denotes here is research-logic synthesis, not the act of writing the result down as prose. Whether the logic stands up when checked against the types and how it is unfolded into readable prose are separate processes, and the model and the support mechanism of this paper address the former. Existing support methods stop short of this problem. AI writing assistance and prompt collections shorten working time yet leave AI positioned as a capable servant (Cui et al., 2024), while support for reflection and for self-regulated learning entrusts the criterion of evaluation to subjective "noticing"

(Schön, 1983; Zimmerman, 2000) and therefore cannot operate, structurally, for a user to whom the discrepancy is not visible in the first place. Prompting a learner who rates their own posed problem as "fairly difficult" to look back detects no discrepancy, because there is no yardstick against which to compare. Learning-by-problem-posing support systems treat constraint violations as errors and converge on the correct answer (Hirashima et al., 2008), and scaffolding grounded in cognitive load theory holds no perspective from which load could be designed as an educational resource (Sweller, 1988; Wood et al., 1976). Proposals calling for metacognitive support and for the cultivation of evaluative judgement (Bearman et al., 2024; Tankelevitch et al., 2024) likewise present design requirements without descending to the level of mechanism. What all of them lack in common is a mechanism that mechanically detects the discrepancy between subjective construction and objective structure and recirculates that discrepancy itself into the next act of construction as the driving force of revision.

1.3. The Core of This Paper

What this paper proposes is a dual-layer model of S&A reciprocity together with Vibe Compiler, the support system that implements it. The claims that form its core come down to the following three.

> (A) The origin of the structural gap. The dissonance that stimulates metacognition can be classified into four types according to whose Synthesis and whose Analysis it arises between. What this paper adopts is the fourth type, in which the AI's Analysis probes the results of the AI's Synthesis and thereby excites the human's metacognition (Section 3, Section 3.2).
>
> (B) Articulation through a twofold distinction. The four quadrants obtained by superimposing the distinction between the executing agents, human and AI, upon the functional distinction between Synthesis and Analysis are what make (A) possible. Without this distinction, what the AI evaluated and what the user was able to evaluate, and what the AI made and what the user constructed, collapse into the single word "done" (Section 3, Section 3.3).
>
> (C) What drives generative AI is content. The prototype is constituted not by prompt engineering but solely by feeding in seven kinds of materials. Ontological documents written for human readers, never having passed through formalization, ran on the LLM as the skeleton of inference exactly as they were. The centre of gravity of value has moved from the skill of formalizing structure toward the content, toward the question of what ought to be given structure (Section 4, Section 4.2; Section 6, Section 6.1).

What is placed at the foundation of all of these is S&A reciprocity. This paper decomposes intellectual construction into Synthesis, which selects and combines known logical parts to suit a purpose, and Analysis, which objectively maps the constructed artifact against the structural complexity specific to the domain, and it regards the mutual traffic between the two as the source of learning and of logic generation. The key

lies in a structure of mutual constraint, in which the output of Analysis is not consumed as an evaluation result but immediately flows back as a constraint on the next Synthesis. It is this recirculation that makes reciprocity a dynamic mechanism of construction rather than an activity of evaluation.

On this foundation, the paper actively exploits the dissonance between what the user intended and the objective structural indicators the AI presents, taking it not as an obstacle to be removed but as a source of metacognitive stimulation. We call this structural-gap-driven metacognitive support. The role of AI is redefined here not as a capable servant that supplies answers but as a critical file, in the sense of the rasping tool, which deliberately probes the fragility of premises and the blanks in the logic. By withholding answers it maintains productive struggle, and the user is left with no choice but to decide the direction of revision themselves. Continually handing back the authority to decide and the accountability that goes with it is what drives the elevation from maker to manager. This redefinition is a requirement rather than an option, because a large language model left to itself will side with the user (Sharma et al., 2023), and the role of a critical partner does not come about without explicit design.

The model forms a dual-layer structure. The first layer takes the learner's metacognition as its object and cultivates evaluative judgement in arithmetic problem posing and in reading comprehension in language arts. The second layer takes the researcher's metacognition as its object and feeds the support logic itself, the question of how the learner's metacognition is to be stimulated, into a type check against a paper ontology. The two layers share one and the same reciprocity mechanism, and what is swapped out is only the content of the Analysis mapping. This paper itself is an output of the second layer, so that a relation of self-application holds.

1.4. Contributions

The reciprocating structure of refining an artifact through repeated generation and evaluation is not itself an invention of this paper. Isomorphic cycles are already established in the Geneplore model (Finke et al., 1992), in the co-evolution of problem and solution (Dorst & Cross, 2001), in the Analysis–Synthesis Bridge (Dubberly et al., 2008), in the iterative cycle of engineering design (Asimow, 1962), and in the reciprocity of design and analysis in design-based research (Brown, 1992). This paper acknowledges the fact frankly. That the reciprocating structure is pre-existing and that every model employing that structure is the same are, however, entirely different matters. When a vocabulary carries a definite meaning, when that meaning implies a concrete operation, and when that operation is differentiated from others, novelty is constituted. Driving an engine that stimulates metacognitive function through the distinction between S and A is not explained by the fact that the reciprocating structure is shared. What the existing reciprocity models have described is a cognitive process that arises spontaneously in the expert, and neither the case in which reciprocity fails to occur nor the question of who carried each side of the reciprocity can be posed within them. The contribution of this paper lies on the side of the mechanism by which reciprocity is excited from outside.

1. A dual-layer model of S&A reciprocity. We formalize the mutual-constraint loop

that recirculates the output of Analysis as a constraint on Synthesis, and show that the learner layer and the researcher layer are supported by one and the same mechanism with only the Analysis mapping swapped out.

2. Four types of the origin of the structural gap. We supply a vocabulary that describes the design intent of a support system in the GenAI era in the form of whose Synthesis or Analysis the gap is excited against.
3. An articulation of the roles of human and AI through four quadrants, which makes it possible to describe as distinct phenomena what had until now collapsed into the single word "done."
4. The design and prototyping of Vibe Compiler. We operate the 16 parameters of the paper ontology as a type system, and give concrete shape to a dialogue design and a user interface equipped with Null checks, consistency checks, probing triggers, and an acceptance path for reverse Analysis.
5. The finding that content drives generative AI. We show, through the actual configuration and the execution logs, that feeding in unformalized, content-oriented structured documents can make generative AI function as a compiler of research logic.

1.5. Assumptions and the Scope of the Claims

The proposal of this paper holds on four assumptions: that the objective structural indicators the AI computes are obtained at an accuracy acceptable in practice; that the system can present a reference solution for a given artifact (AI as Oracle); that parameters expressing structural complexity can be defined by hand for each target domain; and that the user possesses a minimum of domain knowledge and can respond to the AI's remarks with grounds. The first two can break down as difficulty rises, but this paper builds that breakdown into the design not as a defect but as an occasion for evaluation that elicits the user's counterargument, because a grounded counterargument against the AI's evaluation (reverse Analysis) is precisely the observation point for whether epistemic agency remains on the human side. The last assumption places application to complete novices outside the present scope.

We draw a sharp line between what has been verified and what is stated as a prediction grounded in the design. That the prototype executed the type check against the paper ontology and actually built the logical structure of this paper is a fact supported by the execution logs (Section 4). Whether presenting the structural gap improves learners' evaluative judgement to a statistically significant degree, by contrast, is a matter calling for quantitative evaluation under controlled conditions, and this paper holds no such data. Five research questions can be formulated under the model: whether the type check detects Null slots and brings the user to articulate them (RQ1); whether presenting the structural gap converges the prediction error E_{pred} and widens the evaluation coverage S_{cov} (RQ2); whether a design that returns questions maintains productive struggle (RQ3); whether the dual-layer structure holds across layers and domains (RQ4); and whether reverse Analysis functions as the indicator of epistemic agency A_{epi} (RQ5). This paper gives positive evidence for RQ1 and RQ4, and leaves RQ2, RQ3, and RQ5, which ask

after the magnitude of the effect, at the level of formulation, deferring their verification to a separate paper.

The remainder of the paper is organized as follows. Section 2 establishes what distinguishes this work from related research, and Section 3 formalizes the proposed model. Section 4 presents the specification and the results of the prototype, Section 5 shows examples of its application to both layers, Section 6 discusses the implications and the limitations, and Section 7 draws the argument together. An appendix records the prehistory out of which the prototype came about.

2. Related work and the position of this paper

The role of this section is to state critically how the model proposed here differs from each of the existing lineages, and thereby to build the logic of that difference. What has to be acknowledged at the outset, ahead of any defence, is that the S&A reciprocity proposed here shares its loop structure with the existing family of reciprocity models. The substantive work of this section is to identify, granting that shared structure, the point at which the present model diverges.

2.1. The fields this paper connects to, and what they have left unresolved

What the fields this paper connects to have left unresolved takes a common shape. Each is equipped with a vocabulary for asking what is happening on the human side, yet lacks a mechanism that renders that question observable, and so none arrives at the position from which one can ask whether metacognition remains on the human side at all.

Learning by problem posing is a domain that Silver (1994) formalised as an activity carried out before, during, and after problem solving, and to which Christou et al. (2005) gave a taxonomy of editing, selecting, comprehending, and translating (for the overall picture see Cai et al., 2015; 2023). What has accumulated there is product-centred knowledge, with competence as Makers in its sights. What has been left unresolved is the evaluative process, that is, by which measure the poser judged their own construction, and, further back still, an account of the very mechanism by which problem posing promotes learning (Cai & Hwang, 2020). The S&A reciprocity responds to this residue directly, by explaining problem posing as an activity that forces a reciprocity between Synthesis and Analysis, and learning as something that occurs because that reciprocity stimulates metacognition. The same void appears in research on metacognition and evaluative judgement. Flavell (1979) defined metacognitive monitoring, Zimmerman (2000) supplied a model of self-regulated learning (for a comparison see Panadero, 2017), Tai et al. (2018) established the concept of evaluative judgement, and Bearman et al. (2024) argued for the necessity of cultivating it in the age of generative AI (on the perspective of sustainable assessment see Ajjawi et al., 2018). Self-regulated learning, however, leaves the criterion of evaluation inside the learner, and evaluative judgement remains a normative claim that it ought to be cultivated. In neither is there a mechanism that forces objective structural indicators and subjective self-assessment into confrontation. It is in

order to fill this void that the present paper defines the structural gap quantitatively and operationalises the improvement of evaluative judgement as the convergence of E_{pred}.

Matters run in parallel in the lineage concerned with agency. Research on epistemic agency, set against the background of the theory of human agency of Bandura (2006), has knowledge building theory (Scardamalia & Bereiter, 2014), collective cognitive responsibility (Zhang et al., 2009), and shared epistemic agency (Damşa et al., 2010) at its core, and has shown that agency emerges not as a static capability but as a process of negotiation and redistribution (Stroupe, 2014). These findings give theoretical warrant for treating agency as an observable indicator A_{epi}, but because what they treat is negotiation among human beings, the case in which the counterpart is an artefact capable of taking over one whole side of intellectual production is not envisaged. Without a vocabulary for asking who executed what, a state of affairs in which agency is being handed over cannot be distinguished from one in which a result was achieved jointly. The four quadrants introduced here are introduced precisely in order to supply that distinction. Discussions of agency in the age of generative AI fill part of this void: Cox (2024) presents three views of educational purpose, namely Makers, Managers, and information organisms, and together with the blurring of responsibility that accompanies delegation (Dwivedi et al., 2023), the theory and evidence of cognitive offloading (Gerlich, 2025; Kosmyna et al., 2025; Lee et al., 2025; Risko & Gilbert, 2016), and the foundations of productive struggle (Hiebert & Grouws, 2007; Warshauer, 2015), this body of work gives an outline of what is damaged when AI is put to work as a capable servant (for learners' acceptance see Chan & Hu, 2023). What is left unresolved is the level at which one asks by which indicator the agency of the human side is to be observed and by which mechanism it is to be recovered. Inheriting Cox's normative framework, this paper brings it down to a question of mechanism: what kind of dialogue design actually brings about the transition from Makers to Managers.

The two lineages on the design side stop at the same place. What hybrid intelligence (Akata et al., 2020; Dellermann et al., 2019) asks is how to optimise the division of labour, and under an objective function of maximising the outcome no term can be defined that expresses the benefit of deliberately not letting a task be taken over (that the combination of human and AI does not always surpass either alone has been shown by Vaccaro et al. (2024)). Designating Q3 as the quadrant to be reserved for the human side is the operation that introduces such a term explicitly. Critical digital pedagogy (CDP) (Morris & Stommel, 2018; Selwyn, 2019; Stommel et al., 2020) supplies scepticism toward designs that treat efficiency as an unconditional good, and affords a standpoint from which to examine digital poverty (Prather et al., 2024) and the monoculturing of knowledge (Messeri & Crockett, 2024), but what it supplies is a critical standpoint rather than a constructive theory of design. Critique points at where the problem lies; it does not give a mechanism. And what Tankelevitch et al. (2024) and Bearman et al. (2024), the works closest to the present one, put forward was likewise a design requirement, that metacognitive support is needed. A requirement points at where support is to be placed, but has no words for measuring what remains once it has been placed. By introducing the structural gap as an observation point and the four indicators as measures, this paper

brings that criterion down to the level of mechanism.

2.2. Differences from existing learning support methods

The established standard methods share a design philosophy. Scaffolding withdraws external support by degrees and transfers responsibility (van de Pol et al., 2010; Vygotsky, 1978; Wood et al., 1976), reflection prompts ask after the activity what the learner has noticed (Boud et al., 1985; Schön, 1983), and rubrics present the points of evaluation in advance and have learners score themselves (Panadero, 2017). Automated feedback, as in the problem-posing learning support system MONSAKUN, judges the output from the standpoint of constraint satisfaction and points out errors (Hirashima et al., 2008; 2014; Supianto et al., 2017), and the standard form of generative AI literacy education is likewise an application of self-regulated learning (Anders & Dux Speltz, 2025); from the standpoint of cognitive load theory all of these have been justified as the removal of extraneous load (Kirschner et al., 2006; Sweller, 1988). What they have in common is that the criteria of evaluation are given from outside while the difference between those criteria and the learner's subjective sense is never made visible, that support consists in pointing out errors or guiding toward the correct answer rather than in a design that sustains struggle, and that the result of evaluation does not flow back as a constraint on the next act of construction. The model proposed here stands at the inverse of these three points. It is less that the standard methods have failed to solve them than that, under a design whose purpose is to remove extraneous load and guide toward the correct answer, they never arise as goals in the first place.

The relation that calls for the most careful distinction is that with reflection and self-regulated learning. The S&A reciprocity resembles the cycle of self-regulated learning (Zimmerman, 2000) in outward form, but what corresponds to Analysis there is introspection internal to the learner, and the relation in which an objective indicator intervenes from outside and recirculates as a constraint on the next Synthesis is not thematised. Moreover, for a user who cannot see the discrepancy between the subjective and the structural in the first place, looking back does not operate at all, and this for structural reasons. Organising the difference this paper claims into three, then, it consists in the objective externalisation of the indicators, whereby a measure in the form of quantitative parameters computed by AI is presented and forced into confrontation with the subjective; in the mutual delimitation of construction and evaluation, whereby the result of Analysis immediately recirculates as a constraining condition on the next Synthesis; and in the deliberate maintenance of productive struggle, whereby AI refuses to be a capable servant that supplies the answer and, by confronting the user with the gap between expectation and reality, sustains the struggle of autonomous revision. The greatest difference from existing support systems whose purpose is efficiency lies in this third point.

The difference from existing prompt collections can be explained along the same axis. Most prompt collections treat AI as a capable servant and aim to have the work completed on the user's behalf; their criterion of evaluation reduces to whether the output matches the user's intuition, and their goal is the generation of the output itself. The

approach proposed here, in contrast, defines AI as a critical file, in the sense of the rasping tool, maps the artefact by means of objective parameters, and, by making the gap visible, elevates the user into a Manager. What decisively separates the two is a difference of objective function, between aiming at the shortening of work through efficiency and aiming at the improvement of the human side's metacognitive capacity through dissonance with AI. It should be added that the properties of large language models, which will accommodate the user if left to themselves (Sharma et al., 2023) and whose output carries no guarantee of factuality (Huang et al., 2025), are among the conditions that make the problem addressed here a well-posed one. Constituting AI as a critical file by forbidding it to present revisions and constraining it to return questions is at once a response to the warning of CDP and a member of the family of hybrid intelligence designs in which the human side retains the authority to decide (Damşa et al., 2010; Dellermann et al., 2019).

2.3. Relation to existing models that treat reciprocity

Table 1 organises the relation to the existing family of cognitive and design models that describe reciprocity itself.

The closest of these to the present work is the Geneplore model of Finke et al. (1992). Its generative process, which produces preinventive structures in rough outline, corresponds almost exactly to what is here called Synthesis, and its exploratory process, which examines and interprets those structures and adjusts the constraints on generation, corresponds to Analysis; the implication that the artefact is honed with every iteration is shared as well. As a loop structure, then, the S&A reciprocity is essentially isomorphic to the Geneplore model. What this paper claims as novel lies not in the discovery of this loop but on the side of the mechanism that excites it from outside, under conditions in which it does not arise of its own accord. That the question raised here cannot be raised inside that model despite the isomorphism follows from the fact that Geneplore is a theory describing, after the fact, cognitive processes that arise spontaneously within a creative individual, and contains neither a procedure for deliberately exciting the reciprocity nor a mechanism for naming from outside which of the two processes the executor is currently in. In a theory that presupposes a single executor, asking who the executing agent is carries no meaning. What this paper asks lies outside that presupposition: what remains on the human side once AI has taken over one side of the reciprocity, and how a mechanism that artificially excites the reciprocity under those conditions is to be designed. It is precisely because the structures are isomorphic that it becomes clear that this paper's question cannot be raised inside the existing model. The other five models stand in the same position. The co-evolution of Dorst and Cross (2001) anticipates the mutual delimitation described here in observations of experts, but does not treat the case in which co-evolution fails to occur, and the bridge of Dubberly et al. (2008), the process design of Asimow (1962), and the DBR of Brown (1992) and colleagues (Collins et al., 2004; The Design-Based Research Collective, 2003) describe transformation or cycling while leaving untouched the level at which one asks who performs it and whether the executor is aware of doing so. The single exception is Sowden et al. (2015), which comes

Table 1: Transition triggers and the treatment of awareness of mode in existing models that treat reciprocity

Model (representative reference)	Counterpart of Synthesis / counterpart of Analysis	Transition trigger	Awareness of mode (metacognition)
Geneplore (Finke et al., 1992)	Generative process / exploratory process	Implicit judgement internal to the executor	Not thematised
Co-evolution model (Dorst & Cross, 2001)	Proposal of a solution / understanding of the problem	Change in the perception of the problem brought about by producing a solution (internal)	Not thematised
Analysis–Synthesis Bridge (Dubberly et al., 2008)	what could be / what is	Construction of an abstract model (the bridge) (internal)	Not thematised
Dual process and Shift (Sowden et al., 2015)	Divergent thinking / convergent and critical thinking	Metacognitive control by the executor (internal)	Treated as individual differences in the capacity to Shift (external support not treated)
Analysis–Synthesis–Evaluation (Asimow, 1962)	Construction of a solution / decomposition of requirements and judgement of their satisfaction	Prescribed as a procedural norm	Not thematised
DBR (Brown, 1992; The Design-Based Research Collective, 2003)	Design of the learning environment / analysis of practice	Malfunctions observed in practice (the researcher's interpretation)	Not thematised
This paper (S&A reciprocity)	Construction / structural evaluation	The gap between subjective self-assessment and objective structural indicators (an externalised, observable trigger)	AI names the mode and what is missing, and prompts the user toward awareness

closest to lending support to the present argument in holding that the capacity to move back and forth between the two modes (Shift) and its metacognitive control govern the quality of the artefact; what it treats, however, is individual differences in Shift and their mechanism, and not a state in which awareness itself has been lost, nor one in which an external agent performs one side of the switching on the executor's behalf (on the origins of the dual-process account see Cross, 2006; Guilford, 1967).

What the table shows to be unthematised converges on two matters, namely where the occasion for transition resides and awareness of mode. In the existing models, the transition between modes has been explained either as a tacit sense of unease arising within the executor or as a procedural norm. So long as it remains internal, whether a transition is occurring cannot be asked from outside, and for that reason cannot be excited from outside either. Only once the structural gap has been put in place as a vocabulary does this question take an observable form and become, at the same time, an operable object. The treatment of cognitive load runs likewise: so long as cognitive load theory (Sweller, 1988) positions load as a quantity to be removed, the demand to leave load in place deliberately cannot appear as a goal that can even be described. The difference from the "minimal guidance" criticised by Kirschner et al. (2006) lies in the fact that the AI envisaged here does not stand back but points out concretely the unfilled slots and the gap; what is at issue is not the amount of support but what remains on the human side under conditions in which support continues to exist.

2.4. The lineage of content-oriented ontology engineering

Finally, as the lineage to which this paper is most deeply indebted, we take up the claim of content-orientation in ontology engineering. This lineage occupies a place unlike the others in that it gives the framework that explains why the Vibe Compiler worked.

Bourdeau and Mizoguchi (2000) stated explicitly that the difficulties obstructing the construction of intelligent educational systems are all problems that concern content, or in other words that neither inference technology nor beautiful theoretical formalisation contributes to improving the situation. The same claim is restated in Bourdeau and Mizoguchi (2016) in the form that the distinctions an ontology draws are not a matter of the form of representation on a computer. What ontology engineering has consistently regarded as important is not the refinement of expression in a formal language but the content side, that is, what to posit as concepts and what constraints to place among them (Mizoguchi, 2004). This claim is, if anything, reaffirmed all the more strongly in the age of generative AI. However powerful an LLM becomes, unless a structure is given to it, fluent prose is all that is produced and no dissonance arises. What was actually doing the work in our prototype was not the inference engine but the structure of the content that had been fed in, and moreover that paper ontology was not a heavy-weight formal description but a prose document written with human readers in mind. The fact that it ran as the skeleton of inference on an LLM without having passed through formalisation carries a new implication for this lineage. We position this not as a competitor to ontology research but as a spiritual sequel that raises the importance of content-orientation anew, from the aspect of protecting agency (see Sections 6 and 6.1).

From all of this the position of the present paper is fixed as follows. This paper is not the discoverer of the structure called S&A reciprocity. What it presents is a design that becomes possible only by superimposing the distinction of executing agent on this already known structure, that is, a mechanism that protects and at the same time drives the metacognitive function of the human side under conditions in which AI can take on one side of the reciprocity. The existing family of reciprocity models has no vocabulary for asking who executes each edge of the reciprocity, and DBR treats a division of labour among multiple human agents but does not envisage a state of affairs in which an artefact takes on the Analysis edge. Content-oriented ontology engineering, for its part, has gone on asking what ought to be made into structure, but did not envisage the conditions under which that structure runs as an inference engine without passing through formalisation. This paper stands at the intersection of these.

3. The Proposed Model: Structural-Gap-Driven Metacognitive Support

3.1. The Underlying Idea

At the base of the model lies the dissonance between the subjective intention a user holds toward an artefact and the structural reality of that artefact as computed in the form of objective indicators. This dissonance is a source that stimulates metacognition (Flavell, 1979); it is not a defect to be removed but the occasion that makes the user ask anew what, and in what way, they had misjudged.

The S&A reciprocity is the core concept that carries this idea. This paper decomposes intellectual construction into Synthesis, which selects and combines known logical components according to a purpose, and Analysis, which critically evaluates the artefact against objective criteria, and it locates the source of learning effects and of logic generation not in a one-directional linkage of the two but in their mutual interchange. Analysis consists of three layers: the verification of logical consistency, which asks whether the artefact functions without breakdown; structural and metacognitive evaluation, in which the user asks against quantitative parameters whether the artefact carries the structural complexity that was intended; and value-oriented, inquiry-oriented evaluation, which asks whether the artefact is interesting and whether it strikes at the essence of the original purpose. What matters is that the output of Analysis is not consumed as an evaluation result but flows back immediately as a constraint on the next Synthesis. This structure of mutual constraint is what turns the S&A reciprocity from an evaluative activity into a dynamic mechanism of construction.

Underlying this idea is a reinterpretation of the role of AI. Left unchecked, AI ingratiates itself with the user (Sharma et al., 2023) and behaves as the capable servant (Cox, 2024) that hands over whatever answer is demanded. This invites cognitive offloading (Risko & Gilbert, 2016) and erases the productive struggle (Hiebert & Grouws, 2007; Warshauer, 2015) that is the route to deep understanding. What this paper sets against it is a design that constitutes AI as "a critical file, in the sense of the rasping tool", one that deliberately abrades the fragility of premises and the blanks in a line of reasoning. The file offers no revisions; it presents objective indicators and then responds only in the

form of questions. This constraint forms the condition under which evaluative judgement is honed while the struggle is sustained.

3.2. Where the Structural Gap Is Born: Four Types

This subsection is the core of the model. Although the loop structure of the S&A reciprocity is itself not new, the point at which the model proposed here diverges decisively from existing models is the question of the origin of the structural gap: between whose Synthesis and whose Analysis is that metacognition-stimulating dissonance born?

The four types come into view once the single scene of research activity is laid out side by side (Figure 1). What divides them is not whether a gap exists but which quadrant it originates from, and once the origin moves, the object that is called into question moves with it.

(I) Research activity by humans alone. The gap is generated inside the person who carries out the work. Building a line of reasoning from intuition and holding the hunch that it is promising, the researcher criticises their own Synthesis result, and the difference between that hunch and the criticism rises into awareness. What the Geneplore model (Finke et al., 1992) and the co-evolution model (Dorst & Cross, 2001) have described is this internal genesis. The limitation here is not that no gap arises but that blind spots remain, because the critical eye is also one's own. The traditional form of research supervision, in which an advisor or a reviewer returns questions, works as a reinforcement, yet it is irregular and dependent on the individual.

(II) Wholesale delegation to generative AI. Since both Synthesis and Analysis are completed inside the AI, the human side has no origin for a gap. The user only throws in a topic, never voices a hunch, and does no more than skim the returned text and approve it, while on the AI's side generation and evaluation cohere self-justifyingly. Neither the hunch nor the measurement by type is externalised, so nothing appears on either side to be compared, and no counterpart remains at hand against which the person might hold their work up. What must be noticed is that the type of the paper is nevertheless filled in. The authors' draft V1 is an instance of exactly this: the type was almost fully satisfied while the human content stayed empty. Cognitive offloading is a problem not because the work decreases but because human judgement ceases to be needed anywhere.

(III) Existing human–AI collaboration models. This is the route that generates the gap out of an Analysis performed by human and AI side by side, and many designs in hybrid intelligence (Akata et al., 2020; Dellermann et al., 2019) belong here (that a combination does not always outperform either party alone has been shown by Vaccaro et al. (2024)). The human builds the logic, the AI assists as far as drafting without touching the synthesis of the logic, and the two hand down their evaluations separately. What stands here is a divergence between evaluations, so a difference does indeed arise, but its origin lies on the Analysis side. What is called into question therefore reaches only as far as whether one's own evaluation was on target, and it does not reach the thing one was trying to make.

(IV) Vibe Compiling (this paper). The origin is placed on the AI's Synthesis (and Analysis).The user speaks the Vibe together with the hunch that it is promising, and the

AI turns it into a concrete form and synthesises the logic. The AI may generate artefacts and evaluations actively, and indeed should. Its output, however, is presented not to be received as it stands but as a target for the human to scrutinise critically. The crux lies in the single point that, at the moment the AI turns a vague intuition into a concrete object, the difference between the hunch and the concretised shape appears. What is asked of the user is to adjudicate where that difference comes from, that is, to decide whether their own intuition was lax or whether the mapping onto the type was off the mark, and the check that maps onto the type and detects unfilled slots is placed there for the sake of that judgement. What is called into question here is not the evaluation but the thing one was trying to make. The gap is therefore generated deliberately, by design, neither inside the human nor inside the AI but between the AI's output and the human's judgement (Table 2).

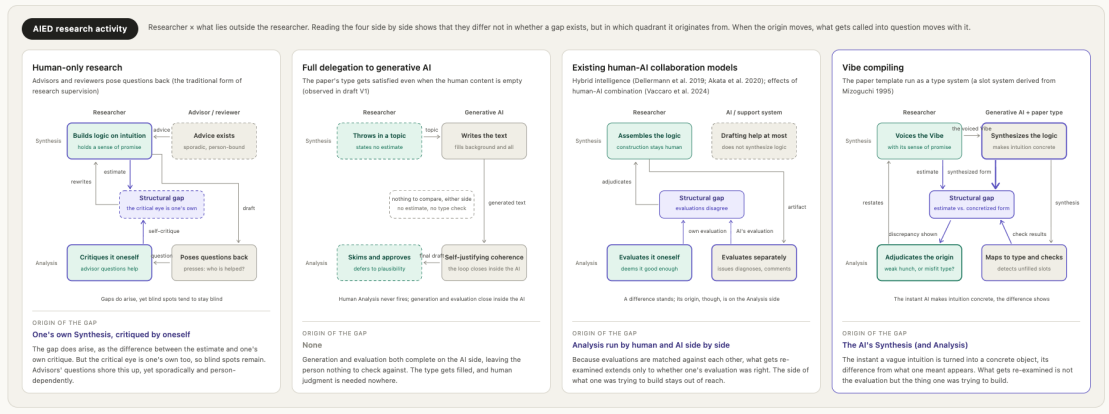

Figure 1: The four types of origin of the structural gap in research activity. Laid out side by side, they make it visible that what divides them is not the presence or absence of a gap but which quadrant it originates from. Once the origin moves, the object that is called into question moves with it.

Only by laying the four types side by side does the following become sayable. What generative AI has made possible is not the automation of Analysis; it is a gap whose origin is Synthesis. As long as the gap arises from the collation of evaluations, what is called into question reaches only as far as whether one's own evaluation was on target. Only when the AI converts intuition into a concrete object does what one was trying to make become an object of questioning.

A note is in order here on how the agent columns of Table 2 are to be read. What an agent column names is the side that principally carries each mapping, not an exclusive right to execute it. In the actual operation of Type (IV), both Synthesis and Analysis are carried by human and AI alike. The principal of Synthesis is the AI: the human gives the Vibe and, on seeing the output, returns supplements and corrections to it. The principal of Analysis is likewise the AI: the human receives the stimulus of a probe, adjudicates whether it holds, and moves toward correcting the result of Synthesis. This correction, that is, the act of supplementing the Vibe or casting a new Vibe, is counted on the side of Synthesis, as an act that produces something new in response to an evaluation.

Table 2: The four types classified by the origin of the structural gap

Type	Agent of Synthesis	Agent of Analysis	Origin of the gap and its limitation
(I) Research by humans alone	Human	Human	Generated internally by self-criticising one's own Synthesis result. Because the critical eye is also one's own, blind spots remain, and the reinforcement supplied by an advisor's questions is irregular and dependent on the individual
(II) Wholesale delegation to generative AI	AI	AI	No origin on the human side. The type of the paper is filled in, but human judgement is needed nowhere (observed in draft V1)
(III) Existing human–AI collaboration models	Human	Human + AI	Generated from an Analysis performed by human and AI side by side. What stands is a divergence between evaluations, and the only available reading attributes the discrepancy to one's own evaluation, so what is called into question reaches only as far as whether one's own evaluation was on target. A route that has the user adjudicate the attribution of the discrepancy could reach the thing one was trying to make as well, but this type has no such route (Section 5.3)
(IV) Vibe Compiling (this paper)	AI + human (chiefly the AI)	AI + human (chiefly the AI)	Generated, at the moment the AI turns intuition into a concrete object, as the difference between the hunch and the concretised shape. The human, taking the AI's output as target, adjudicates whether each probe hits home. What is called into question is not the evaluation but the thing one was trying to make

Were it set apart as a third function, the decomposition into S&A would itself collapse. The reading of each frame in Figure 1 follows the same organisation: the structural gap at the centre represents the structure, the four boxes represent the four operations formed by the two functions Synthesis and Analysis and the two agents human and AI, and the arrows represent relations of influence among the operations. The process of thought that unfolds inside the human after the stimulus of a probe does not appear in these classificatory variables; it is carried not by the description of the types but by the description of the reciprocity (Section 3.5).

These four types therefore function not as a mere classification but as a vocabulary for designing and evaluating research- and learning-support systems in the age of GenAI, because for any system one can now ask against whose Synthesis or whose Analysis that system has the function of exciting a structural gap. A tool in which the AI generates text and the user merely approves it lies in Type (II) and has no gap-exciting function. A tool that scores automatically and returns the result implements part of Type (III), but if it lacks a route for rebutting the scoring result, evaluative judgement is delegated after all. Using this vocabulary, the paper proposes describing the design intent of each

system in the form of a single question: by what function, and against whose mapping, and which of the two mappings, does it seek to excite a gap? Being able to name the site of metacognitive excitation in this way is the first benefit of the model.

3.3. The Four Quadrants: A Double Distinction of Function and Executing Agent

What makes the four types of the preceding subsection possible is the operation of superimposing the distinction between the executing agents, human and AI, upon the distinction between the functions, Synthesis and Analysis. The four quadrants set this double distinction out explicitly, so that every scene of the reciprocity falls into Q1 (human Synthesis), Q2 (AI Synthesis), Q3 (human Analysis), or Q4 (AI Analysis) (Table 3).

Table 3: The four quadrants formed by Synthesis/Analysis and the executing agent (human/AI)

	Executed by the human	Executed by the AI
Synthesis	Q1 (human Synthesis) *Definition*: the act in which the user selects and combines logical components in the light of a purpose and composes an artefact. It includes deciding what to make and what to ask. *Treatment*: the work of combining components may be delegated, but the part that concerns the setting of purpose and value is reserved to the human side.	Q2 (AI Synthesis) *Definition*: the act in which the AI generates the artefact itself (a posed problem, a draft, an answer, a line of reasoning). Efficiency is maximised, but no history of the composition remains on the user's side. *Treatment*: may be used actively, though the load on Q3 grows in proportion to what is delegated.
Analysis	Q3 (human Analysis) *Definition*: the act in which the user evaluates, criticises, or rebuts their own artefact or the AI's against criteria. It includes declaring a subjective evaluation, performing reverse Analysis on the AI's evaluation, and deciding whether to accept the AI's output. *Treatment*: must be reserved to the human side. A_{epi} is the indicator that measures this residue.	Q4 (AI Analysis) *Definition*: the act in which the AI maps an artefact onto objective structural indicators and names gaps and unfilled slots: type checking, computation of structural parameters, and the generation of remarks in the form of questions. *Treatment*: may be delegated, and indeed should be. This is the quadrant the support mechanism ought to carry.

This distinction is not classification for its own sake. In intellectual activity conducted with generative AI, the single fact that an artefact has been completed renders invisible who did what, and the four quadrants are the minimal apparatus for decomposing that conflation into describable parts. Once the four quadrants are introduced, the design guideline of which quadrants ought to be reserved to the human takes a definite shape. Computing structural complexity by a consistent standard is something humans do poorly, and only through externalisation does the divergence between the subjective and the structural become visible, so Q4 is a domain that ought to be delegated. Nor is it forbidden for the AI to generate artefacts (Q2); Type (IV) rather presupposes active synthesis by the AI. The load on Q3 grows in proportion to what is entrusted, however,

and it is precisely the state of using Q2 while abandoning Q3 that constitutes the use of AI as a capable servant, the form this paper takes as its object of criticism. As for Q1, while the combination of components may be delegated, the setting of what to make and of why it is valuable is inseparable from judgements of validity, and is therefore reserved along with Q3.

Lining up these treatments yields, as it stands, a description of what is being lost when work is entrusted to AI. What can be lost is not the artefact but Q3, the decision as to what counts as valid, and, within Q1, the setting of purpose and value. Entrusting Q2 does not by itself cause the loss; leaving Q3 empty while entrusting Q2 makes the decision itself drop out. When this paper speaks of protecting metacognitive function while driving it, it means this double demand of reserving Q3 and the purpose-setting portion of Q1 to the human and yet making them actually work, and this demand cannot be written down as a single demand unless the distinction of function and the distinction of executing agent are erected together. Only through this distinction do the conflation of what the AI could solve (Q4) with what the learner could evaluate (Q3), the conflation of what the AI made (Q2) with what the learner constructed (Q1), and the state in which the posed problem itself holds up while the self-evaluation departs from the structural reality (Q1 succeeding while Q3 fails) become describable as separate phenomena. Conventionally these have been lumped together under the single phrase "the quality of problem posing", with no vocabulary given for discussing them individually. The paper presents the explanatory power gained through the four quadrants as one of its principal contributions.

3.4. The Dual-Layer Structure: The Learner Layer and the Researcher Layer

This paper develops the S&A reciprocity as a dual-layer structure. The first layer takes the learner's metacognition as its object and, in arithmetic problem posing and in reading comprehension, cultivates the evaluative judgement (Bearman et al., 2024; Tai et al., 2018) that re-evaluates one's own construction by objective indicators. The second layer takes the researcher's metacognition as its object and debugs the researcher's own logical construction by pouring the support logic itself, namely how the learner's metacognition is to be stimulated, into the type check of the paper ontology. Here lies the function that keeps researchers from lapsing into improvements upon rails laid by others and presses them to wring out a novelty of their own. Both layers share the same reciprocity mechanism, and what is exchanged is only the content of the objective parameters. This fact means that the dual-layer structure is not an expedient metaphor but two instances of one and the same model. The prototype that gives the model concrete form on a computer is the Vibe Compiler (Section 4), and since this paper is itself an output of the second layer, a relation of self-application obtains.

3.5. Formalisation

The formalisation below is a descriptive apparatus for designating the constituents of the model without ambiguity. We write the user's purpose as v. The formalisation rests on the premise that the system can present, for any artefact the user submits, a

correct answer together with its solution process (the solvability of the AI, that is, AI as an Oracle), and only under this premise does a reference solution exist against which the learner's own Analysis can be contrasted. The Oracle can nevertheless err, and this paper weaves that error into the design as an opportunity for evaluation rather than as a defect. The act of a user rebutting a fallible Oracle with grounds is the reverse Analysis that A_{epi} of Definition 3 captures, and the Oracle premise and the design of reverse Analysis are two sides of one coin.

Definition 1 (Synthesis mapping, Analysis mapping, and reciprocity). The Synthesis mapping S generates an artefact x by selecting and combining logical components under a purpose v. The Analysis mapping A maps an artefact x onto a domain-specific vector of objective parameters $p = A(x) \in \mathbb{R}^m$. Writing the artefact at the k-th reciprocity as $x^{(k)}$ and the subjective evaluation vector that the user declares in advance at that point as $\hat{p}^{(k)} \in \mathbb{R}^m$, the reciprocity is given by

$$p^{(k)} = A\big(x^{(k)}\big), \qquad x^{(k+1)} = S\big(x^{(k)} \,\big|\, v,\ g^{(k)},\ \mathrm{Null}(\mathcal{O}, x^{(k)})\big) \tag{1}$$

where $g^{(k)}$ is the structural gap of Definition 2 and $\mathrm{Null}(\mathcal{O}, x^{(k)})$ is the set of unfilled slots of Definition 4. The second equation of Eq. (1) is the core of the model. That the output of Analysis appears as an argument of the next Synthesis, that is, that mutual constraint is written down as a structure, is the formal difference from introspective models such as Reflection and self-regulated learning.

Definition 2 (Structural gap and convergence). We define the structural gap $g^{(k)}$ at the k-th reciprocity as the weighted distance between the subjective evaluation vector and the objective parameter vector.

$$g^{(k)} \;=\; \big\|p^{(k)} - \hat{p}^{(k)}\big\|_W \;=\; \left(\sum_{i=1}^{m} w_i \left(p_i^{(k)} - \hat{p}_i^{(k)}\right)^2\right)^{1/2} \tag{2}$$

The $w_i > 0$ are normalisation coefficients that absorb differences of scale across domains. A large $g^{(k)}$ expresses the divergence between what the user intended and the reality of the artefact, and hence a state with ample room for metacognitive stimulation. A sequence of reciprocitys is said to have converged when, for a threshold $\varepsilon > 0$, $g^{(K)} < \varepsilon$ holds and $\mathrm{Null}(\mathcal{O}, x^{(K)}) = \emptyset$. What is valued as a learning outcome, however, is not the smallness of $g^{(K)}$ on individual artefacts but the exhibition of a small initial gap $g^{(0)}$ on novel tasks as well, that is, the internalisation of Analysis, the acquisition of an "eye" that can hit upon the objective indicators unaided, without the AI.

Definition 3 (Four indicators of metacognitive honing). We define the metacognitive honing brought about by the reciprocity as four domain-independent indicators. The parameter prediction accuracy E_{pred} is the mean absolute error normalised by the range $r_i = p_i^{\max} - p_i^{\min}$, and improvement in evaluative judgement is operationalised as its monotone decrease and convergence. The evaluation coverage score S_{cov} is the coverage, by the set M_k of indicators mentioned in the k-th self-evaluation utterance, of the quality

indicator set Q ($|Q| = 5$, items 11–15 of Table 4) consisting of accuracy, efficiency, reliability, scalability, and reusability, and its rise expresses that the viewpoint of evaluation has spread from a single axis to multiple dimensions. The refinement consistency L_{ref} is the resolution rate of the set I_k of shortcomings that the user themselves pointed out in the k-th Analysis, and it distinguishes revision made "somehow" from strategic revision grounded in evaluation.

$$
\begin{aligned}
E_{pred}(k) &= \frac{1}{m}\sum_{i=1}^{m}\frac{\left|\hat{p}_i^{(k)} - p_i^{(k)}\right|}{r_i}, \qquad S_{cov}(k) = \frac{|M_k \cap Q|}{|Q|}, \\
L_{ref}(k) &= \frac{\left|\{\, d \in I_k \;:\; d \text{ is resolved} \,\}\right|}{|I_k|}
\end{aligned}
\tag{3}
$$

The fourth indicator, the epistemic agency score A_{epi}, is defined from the quantity and the quality of the rebuttals with established grounds (reverse Analysis) that are directed at the AI's evaluations. Let E be the set of occasions on which the AI presented an evaluation and $\mathbb{I}[\cdot]$ the indicator function.

$$
A_{epi} \;=\; \frac{1}{|E|}\sum_{e\in E} w(e)\cdot \mathbb{I}\big[\text{the user rebutted } e \text{ with established grounds}\big] \tag{4}
$$

The weight $w(e) \in \{1, 2, 3\}$ expresses the quality of the rebuttal: 1 for an expression of disagreement accompanied by grounds, 2 for the identification of a mismatch in assumptions or in coverage, and 3 when the identification is accompanied by the proposal of alternative parameters or alternative criteria. The essential point is that the counting is not restricted to occasions on which the AI presented a mistaken evaluation. Under such a restriction the user would be given no opportunity to rebut unless the AI erred, and the observation of agency would end up waiting upon the AI's mistakes. Even when the AI's evaluation is itself correct, if the user objects on the grounds of their own assumptions or design intent and those grounds hold, that is counted as a manifestation of epistemic agency. A_{epi} is an operational definition that brings epistemic agency, an abstract concept, down to observable behaviour, and it is also the point at which the incompleteness of AI is reread as an opportunity for observation rather than as a defect. The concrete values of ε, w_i, and $w(e)$, together with the procedure for judging whether the grounds of a rebuttal hold, are design matters that require calibration.

These four indicators function as debugging indicators for the support logic not only for learners but also for the researchers who develop the system. A high A_{epi} on the learner's part is evidence that the critical file the researcher designed is stimulating the learner's agency correctly, and thus indicates the success of the second layer. This double reading of the indicators is a consequence of the dual-layer model.

Definition 4 (Paper ontology, type checking, and Null determination). Let the paper ontology be $\mathcal{O} = \{o_1, \ldots, o_{16}\}$. Each o_j is a mandatory slot constituting academic logic, and their contents are given as the 16 academic parameters of Table 4. We write the

value of slot o in artefact x as $\mathrm{val}(o, x)$, and set $\mathrm{val}(o, x) = \bot$ when it is unfilled.

$$\mathrm{Null}(\mathcal{O}, x) = \left\{\, o \in \mathcal{O} \;:\; \mathrm{val}(o, x) = \bot \,\right\},$$
$$\mathrm{TypeCheck}(x) = \begin{cases} \text{true} & \big(\mathrm{Null}(\mathcal{O}, x) = \emptyset\big) \\ \text{false} & (\text{otherwise}) \end{cases} \tag{5}$$

When $\mathrm{TypeCheck}(x) = \text{false}$, the system fires a compile error for each element of $\mathrm{Null}(\mathcal{O}, x)$. It is decisive that the error comes back not as a proposed revision but as a question that makes the user verbalise the slot in question. Null determination here covers not only the absence of a value but also the case in which a description exists yet fails to correspond logically to the other slots. In this sense TypeCheck is a composition of an existence check and a consistency check, the latter of which is described in Section 4.4.

3.6. Domain Generality

While the p that the Analysis mapping A computes is instantiated domain by domain, the four indicators are higher-order indicators that do not depend on the domain. In the domain of arithmetic problem posing, the number of operation steps N_{step} (the total number of arithmetic operations required for the answer) expresses the depth of computation, the number of unknowns N_{var} expresses the breadth of the structure, and the difficulty of formulation D_{map} (the complexity of mapping a natural-language context onto a mathematical model) expresses the difficulty of mapping. In the domain of reading comprehension, the reference distance S_{ref} (the number of paragraphs from the point of the question to the grounds for the answer) expresses the depth of search, the number of logical links L_{link} (the number of paragraphs that must be integrated to derive the correct answer) expresses the breadth of integration, and the difficulty of lexical substitution V_{map} (the degree to which concrete expressions in the text must be paraphrased into abstract vocabulary) expresses the difficulty of mapping. In the domain of research-logic synthesis, the degree to which the 16 slots of Table 4 are filled and the coverage of the quality indicator set Q play this role.

It is worth noting that the sub-indicators of arithmetic and of reading comprehension both correspond to the same three axes of depth, breadth, and difficulty of mapping. What is exchanged is only the way each axis is measured, and this commonality is precisely the ground on which the model holds across domains. Generalising, the model is applicable to any domain for which three axes can be defined: structural depth D_{depth} (the number of processing steps and the depth of the logical hierarchy), compositional breadth W_{width} (the number of variables, components, and viewpoints treated), and transformation difficulty M_{map} (the complexity of the mapping from required specifications to a concrete implementation). What the model demands is thus only three conditions, namely that the artefact be describable as a combination of multiple logical components, that a function mapping its structural complexity objectively be definable, and that the user be able to declare a self-evaluation in advance, and the point that only the substance of A is exchanged holds for the difference between the layers as well. The

Table 4: The 16 academic parameters in five categories that type checking takes as its object (the paper ontology $\mathcal{O}$)

#	Category	Parameter	Academic definition and role (type-check item)
1	(1) Significance and purpose	Significance	Why the problem needs to be solved, in the light of the characteristics of the domain
2		Beneficiary	Identification of those who obtain a direct benefit from the solution
3		Benefit	The concrete and novel value obtained through the solution
4	(2) Assumptions and boundaries	Assumptions	The foundation on which the method holds, such as the reliability of the data and the completeness of the theory
5		Coverage	The scope the research treats, with what is out of scope stated explicitly
6		Technical Requirements	The minimum functional specifications and constraints needed to achieve the purpose
7	(3) Difference and critical comparison	Difference from existing methods	The decisive structural difference from existing approaches
8		Limitations of existing methods	Critical analysis of why conventional methods are insufficient
9		Novelty	The point of transformation in concept, design philosophy, or algorithm
10	(4) Evaluation and quality	Functionality	Whether the functions claimed operate correctly as designed
11		Accuracy	The correctness and validity of the output
12		Efficiency	The acceptability of execution time, computational cost, and memory consumption
13		Reliability	Whether operation is stable under noise and special cases
14		Scalability	The capacity to cope with growth in data volume and problem size
15		Reusability	Whether other researchers can use and inherit it in a general-purpose manner
16	(5) Lessons and open issues	Lessons Learned	The abstract lessons extracted through experiment

Note. The numbers in the first column correspond to o_1 through o_{16} of Definition 4. Category (5) is operated so as to require "open issues" as an output sub-slot paired with the lessons learned. Items 11–15 correspond to the quality indicator set Q of Eq. (3).

same logic extends to component-combination intellectual work in general, so that cyclomatic complexity in programming, the chain length of the argument in essay writing, and the coverage of confounding factors in experimental design can each play the role of the objective indicator. Furthermore, the structural fact that the three axes are shared yields a testable prediction, namely that an eye for structure cultivated in one domain transfers to another. The model presents this transfer hypothesis in a form that can be cast into an experimental design (Section 7.2).

3.7. What the Model Makes It Possible to Say

What the model gives is not a claim of effect but a claim of describability. Three points organise what becomes sayable.

Support can be compared by configuration rather than by convenience. The state described as "work gets done with AI, but one cannot say what is happening" is given coordinates. A function that takes over a person's work and a function that prompts a person's judgement become treatable as different things, and the coarse-grained debate over whether to prohibit or to permit the use of AI moves into a design debate over which arrows to connect. The conditions for assembling this configuration are only two: that the hunch can be elicited beforehand, and that the structure can be mapped onto indicators. Any activity that satisfies these two conditions admits the same configuration, and not only problem posing.

What did not happen can be named. Looking at an artefact, one cannot distinguish what has passed through human judgement from what has not. The authors' draft V1 is an instance: the type of the paper was almost fully satisfied with neither a problem-posing system nor an evaluation in place. Who carried what cannot be recovered from the artefact and can only be preserved as a record. Only by superimposing the distinction of executing agent upon the distinction between Synthesis and Analysis can one state that the human's Analysis never once fired. Measurement of effect can measure only what happened. Naming what did not happen requires a different vocabulary.

The origin of the gap can be pointed at as a position. As long as the gap arises from the collation of evaluations, what is called into question reaches only as far as whether one's own evaluation was on target. Only when the AI converts intuition into a concrete object does what one was trying to make become an object of questioning.

The diagnosis of support that does not work, the design of policies for the use of AI, and the formalisation of the hypothesis to be tested next all issue from this describability. And the model is falsifiable: if the same configuration is assembled and yet human judgement does not occur, it is the explanation that is refuted.

3.8. The Overall Structure of the Model

Figure 2 presents the overall structure.

Figure 2 arranges one cycle of the reciprocity horizontally, the dual-layer structure vertically, and the spine that runs through both layers at the right edge. The crux

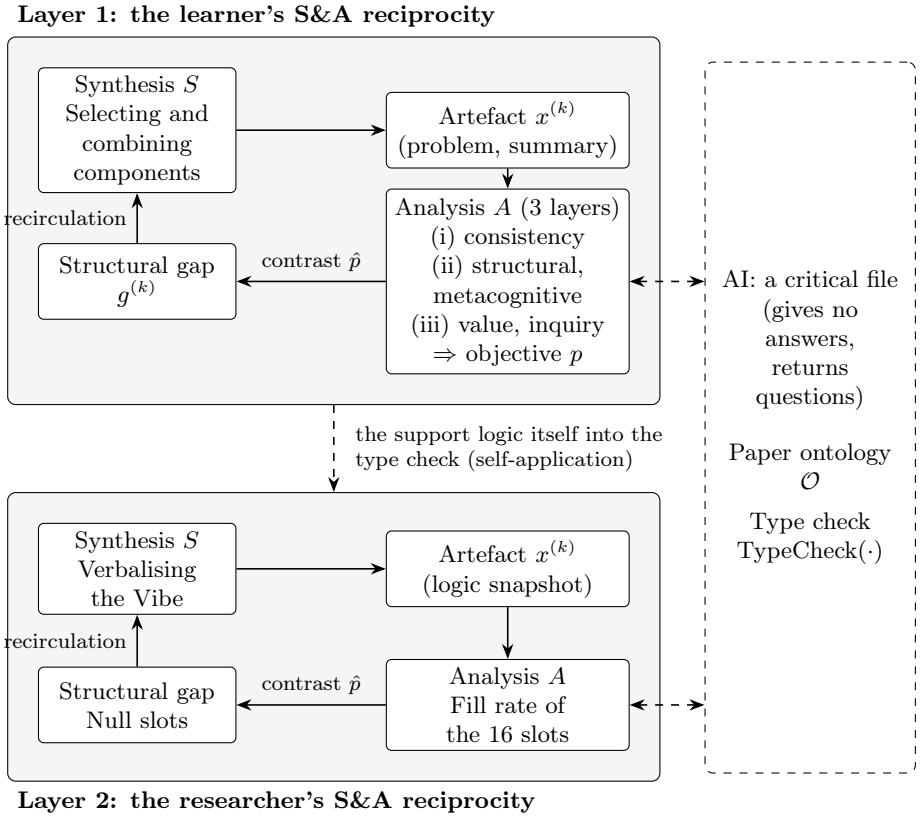

Figure 2: The dual-layer model of structural-gap-driven metacognitive support. Layer 1 (the learner) and Layer 2 (the researcher) share one and the same S&A reciprocity, and AI as a critical file together with type checking by the paper ontology acts as a spine running through both layers.

lies in the recirculation arrow that returns from the lower left to the upper left, which makes visible that the gap and the unfilled slots are given as arguments of Synthesis in the second equation of Eq. (1), that is, that evaluation does not end as evaluation but turns into a constraint on the next construction. Three points should be read off the figure. That the two layers have the same shape means that learner support and researcher support are connected by nothing more than the replacement of the Analysis mapping A. That the spine reaches into the interior of the layers shows that the AI is not a referee handing down evaluations from outside the reciprocity but a constituent built into Analysis (the arrows are bidirectional because the direction from user to AI is the reverse Analysis, whose frequency and quality are captured as A_{epi}). And that the recirculation arrow always returns to the user's Synthesis shows that the authority to decide on revisions remains with the user rather than the AI, which is the design-level guarantee of the redistribution of the four powers (Akata et al., 2020) and of shared agency (Damşa et al., 2010). In terms of the quadrants, the Synthesis nodes are chiefly Q1, the computation of the objective parameters and the detection of Nulls are Q4, and the declaration of $\hat{p}$ together with the decision to accept or reject in response to the

recirculation are Q3. Q2 is not shown explicitly in the figure, not because it is excluded, but because what the figure draws is the main line of the reciprocity; what the critical file forbids is the route that rewrites an artefact without passing through Q3, not generation by AI in general.

4. The Vibe Compiler Prototype: Specification and Outcomes

4.1. Overview and Configuration of the System

The Vibe Compiler is a research-logic compiler that maps a user's inchoate Vibe onto an academic ontology and synthesises it into a logical structure. The system runs a type check on the input and, whenever a constituent that academic writing requires is missing—that is, whenever it is Null—returns feedback in the form of a compile error. What separates this system from ordinary writing-assistance tools is the constraint that what comes back is a question rather than a proposed correction.

The name derives from Vibe coding (Karpathy, 2025). What the AI generates in Vibe coding is working code, and the immediate feedback machinery of a compiler and a runtime guarantees its correctness. In research activity, however, the counterpart of code is not the prose itself but the construction of the logic, and the sequence "background → problem → objective → solution → evaluation → findings" plays the part of the execution log. What this paper proposes is a mechanism that serves as a compiler for that construction of logic; what takes place is therefore not Vibe coding but Vibe compiling.

The prototype is built from NotebookLM and Gemini used in combination, the former carrying the fixity of the "type" that the paper ontology provides and the latter carrying the flexibility of synthesis. What deserves attention is that, in this configuration, the authors did not engineer prompts; they supplied materials. Seven kinds of material were loaded into NotebookLM: (1) a survey paper on agency (Cox, 2024), (2) a paper-template manuscript written by the authors (the paper ontology), (3) the sixteen academic parameters extracted from it (Table 4), (4) the conditions for consistency checks among those parameters, and (5)–(7) three kinds of probing procedure obtained through dialogue with the Vibe Compiler itself (dialogue designs for researchers, for learners, and for the system as a co-creative partner). The survey paper supplies the criteria by which the system can tell a user "you are still at the level of a Maker." The paper ontology is not a description in a formal language but a prose document written for human readers. Each of the three probing procedures was proposed by the system in the course of dialogue with it and then examined and fixed by the authors.

The system's behaviour is governed by the role specification given at start-up, which imposes four conditions: that it define itself as a research-logic synthesis compiler; that it treat the paper ontology as the sole "type" and use the remaining materials as files, in the sense of the rasping tool; that it refuse to accept a mere development report and return unfilled slots as compile errors; and that it respond in the form of questions rather than offering corrections. This is the whole of what was implemented.

4.2. The Core Finding: What Works Is Not the Inference Engine but the Content of the Structure Supplied

The principal finding of this paper follows directly from the configuration just described: what drives the Vibe Compiler is not the inferential power of the large language model but the structure of the content loaded into NotebookLM.

The grounds for this claim lie in a property of the paper ontology that was supplied. That document contains the following passage as a bad example of an introduction: "A system with these functions is required. (No reason is given.) We therefore built such a system. It has these functions. It was built in this way. What is wrong with this is that it is nothing but an enumeration of results and contributes nothing whatsoever to the accumulation of scholarship."

In this example all four items—requirement, system, functions, method—are filled in. A checklist that inspects only whether items are present would therefore pass the document. It fails on one point alone: between the requirement and the system there is no link giving the reason why the system is needed. The document thus preserves, as a test case, a failing example that presence checking cannot detect in principle and that fails only once the relations among the items are examined. The Vibe Compiler can compile because what it was given was not a list of questions but a structure in which conditions hold between one question and another.

This observation amounts to a reconfirmation of what ontological engineering has consistently maintained. When Bourdeau and Mizoguchi (2000) wrote of the difficulties of intelligent educational systems that "all these problems are content-related; in other words, neither inference technology nor beautiful theoretical formalisation contributes to improving the situation," the configuration being pointed at was precisely this one. However powerful an LLM becomes, without a structure to work against it will merely emit fluent prose, and no structural gap will arise. More important still is the fact that the paper ontology supplied had never been formalised. Ordinarily a process of building a formal ontology from such a document would have been necessary. With the advent of generative AI, however, simply loading it into NotebookLM made it run. The implications are discussed in Section 6.1.

4.3. An Execution Example: From Start-up Through Type Check to Error

An excerpt from an execution log at the second layer (the researcher layer) is given below. It comes from a session in which one of the authors submitted their own Vibe concerning learning by problem posing.

> User: You are a system that supports the conduct of research in the manner of Vibe Coding. Proceed with this dialogue as a research-logic synthesis compiler, on the basis of the paper ontology now held in NotebookLM.
>
> System: Understood. Starting as the research-logic synthesis compiler (v1.0). I am a system that polishes and builds your thinking by wielding the other materials as files, so as to fit the fragmentary intuitions (Vibes) you present to the single type that is the paper ontology. This system does not permit a mere development report.

Begin by entering the state of the domain, the problem you seek to solve, and the direction of the solution.

User: Learners ordinarily learn by solving problems, but research on problem posing, which has them learn by composing the problems to be solved, has begun to appear. I want to build a cognitive model that explains why composing problems advances learning. My thought is that decomposing thinking activity into the two components Synthesis and Analysis might make it possible to explain why learning proceeds through problem posing.

System: Performing semantic analysis and mapping onto the types of the paper ontology. The background is the shift from passive problem solving to active learning by problem posing; the problem to be solved is the knowledge gap left by the unexplained internal mechanism behind "why learning advances"; and the proposed method lies in decomposing thinking activity into Synthesis and Analysis and grasping problem posing as their interaction.
[Polishing with the file] According to the materials, a learner's capacity to participate in the construction and evaluation of knowledge is called epistemic agency. Is not the Synthesis and Analysis you speak of precisely the exercise of that agency? Moreover, if Analysis is defined not as mere decomposition but as the evaluative judgement of whether one's own Synthesis is sound, the differentiation from existing cognitive models becomes sharper.
[Type-check error report] The following parameters are missing (Null) from the logic as it stands. (1) Limitations of existing methods: why are existing cognitive models insufficient? Do they overlook S, or A, or the reciprocity of both? (2) Beneficiaries and benefits: who concretely gains what new value? (3) Name of the concept: if you were to give this cognitive model of S&A reciprocity a distinctive name, what image would it convey?

What deserves attention is that although the intuition initially submitted was a single sentence carrying neither a source, nor a beneficiary, nor any differential against existing methods, the system forced it into the slots of the paper ontology and, at the very moment the assignment succeeded, detected what was missing from the chain of logic. The three points raised did not remain isolated remarks: each was converted, through re-synthesis, into settled logic. Point (3) led to the concept name "S&A reciprocity"; point (1) led to the differential logic named "the black-box problem of the cognitive mechanism"; and point (2) connected to "learners and teachers as managers" as beneficiaries and to the benefits they obtain. None of these elements was present in the original utterance. The structural-gap-driven chain of "naming what is missing $\rightarrow$ re-synthesis $\rightarrow$ settling the logic" occurred in an actual dialogue rather than merely as a design assumption. The type check functioned here not as a demand for gap-filling but as the occasion on which new logic was produced.

At each cycle of reciprocity the system externalises the settled logic as a logic snapshot. A snapshot has four fields: the ID of the corresponding element of the paper ontology, the name of the logic, the significance synthesised, and the core of the novelty. Snapshot 01, obtained as the outcome of the session above, gives "S&A reciprocity" as the name of the logic, "decomposing problem-posing activity into Synthesis and Analysis and defining their mutual interchange as the very source of the learning effect" as

its significance, and "incorporating, as Analysis, the evaluative judgement that existing models ignored, thereby making it possible to explain the cognitive mechanism by which problem posing transfers to problem-solving ability" as the core of its novelty. A snapshot is not a static record but an externalisation of the state transitions of the reciprocity process, and it makes traceable which piece of logic was produced in response to which type-check error. Twelve snapshots were finally settled in the process of building this paper's research logic.

4.4. The Mechanism of the Type Check

What the type check returns when it detects an unfilled slot or an inconsistency is laid down as the family of probing triggers in Table 5. Their common purpose is to prevent the cognitive offloading in which an AI robs the user of thought by supplying answers, and thereby to sustain productive struggle. That the error takes the form of a question rather than a proposed correction is the indispensable condition for this.

The type check is not exhausted by checking the existence of slots, because, as Section 4.2 showed, documents exist that fail even though every item is filled in. The Vibe Compiler therefore carries consistency checks that verify the logical correspondence between slots (Table 6). What they verify is whether the academic chain of logic "background → problem → solution → evaluation → findings" is firmly bonded. The mirror-image check between limitations and novelty is of particular note, since it serves to detect the "improvement along rails laid by someone else" into which researchers readily fall: it verifies whether the newness of one's own method is a logical necessity that repairs the weakness of the existing method, and so compels a contrastive structure to be built.

A cross-cutting constraint is further imposed on the triggers in Table 5. Positive synthesis of content that fills a knowledge gap is permitted, and the system may actively synthesise and present the novelty or significance of the research, saying for instance that "that Vibe could become a concrete solution to the lack of scalability from which existing methods suffer." Dissonance and a solution must nevertheless be presented together, and ending an utterance with the criticism alone is forbidden. Anticipatory dialogue that preempts possible consequences is required as well, together with a delegation of choice and control in which several candidate problems are offered for the user to choose among so that a route of rebuttal always stays open, and real-time feedback that confronts the user with evaluation at the moment of construction. These are the conditions under which the AI is recognised as a co-creative partner rather than a judge; what is forbidden is direct rewriting of the artefact, not the presentation of material.

4.5. UI Design: A Logic-Synchronised Development Editor

As the user interface that realises the dialogue design above, this paper proposes a logic-synchronised development editor composed of five panes. The UI is not a matter of mere appearance; it is the apparatus that decides which of the four quadrants the user is placed in. The Main Editor (Q1) is where fragmentary thoughts are written without regard to form, and as they are written the AI attempts, in the background, to map them onto the sixteen parameters. The Sidebar (Q4 and Q3) returns questions in real

Table 5: The system of probing triggers fired by the type check

Layer	Trigger	Example utterance
Common	Slot Null check	"The benefit of the solution is not defined. Describe who is helped and in what way."
Common	Forced extraction of the gap between the ideal and the present state	"Give the domain-specific reason why current technology cannot reach that ideal."
Common	Stress testing against multiple quality criteria	"Has scalability been considered for a hundredfold increase in data volume?" "Are there practical concerns regarding memory consumption?"
Common	Mandatory critical comparison with existing methods	"On what point is your proposal decisively different from conventional methods?" (a claim of novelty that states neither the differential nor the limitations is not accepted)
Common	Prompting the elevation of results into findings	"Organise the current bottleneck as a remaining issue and anticipate the technical elements its resolution will require."
Layer 2	Conversion of intuition into academic parameters	"Is the point that strikes you as unsatisfactory a shortfall in scalability, reliability, accuracy, computational cost, or reusability?"
Layer 2	Probing premises and boundaries	"What premises does that existing method tacitly assume? How is its reliability impaired when they break down?"
Layer 2	Contrast for constructing differential logic	"State the limitation not as a shortfall in performance but as a limitation of the design philosophy. How does your proposal break through it, by relaxing a restriction or by generalising a concept?"
Layer 2	Critical file that stimulates epistemic agency	"Refute the limitation computed by the AI from your own domain knowledge." (the refutation is accepted as reverse Analysis and counted towards A_{epi})
Layer 1	Creation of dissonance through the gap between prediction and result	having the learner declare a self-prediction first, then "The system indicator gives the minimum value. Which element did you find difficult?"
Layer 1	Critical verification of a deliberately lenient evaluation	immediately after issuing a lenient evaluation, "I judged it so, but might I be overlooking a hidden bug or an inefficiency?"

Table 6: Consistency checks among parameters (five kinds)

Name of check	Pair of slots collated	Example of inconsistency
Synchrony of objective and evaluation criteria	quality characteristics contained in the objective ↔ items of the evaluation	the objective proclaims improved reliability, yet the evaluation goes no further than confirming accuracy on small-scale data
Vector collation of significance and benefits	importance and beneficiaries ↔ benefits	the problem is delayed judgement in clinical practice, yet the benefit lies on the unrelated axis of improved system maintainability
Mirror image of limitations and novelty	limitations of existing methods ↔ novelty	the limitation of the existing method is given as high cost, yet the novelty of one's own method is improved accuracy
Boundary between premises and coverage	premises ↔ coverage	the premises assume clean data, yet the coverage includes noisy real-time data
Entailment from experimental results to findings	soundness of the evaluation ↔ findings obtained	the experiment shows that processing is delayed on large-scale data, yet the findings generalise to effectiveness in every environment

time, such as "on what point is your current implementation decisively different from existing methods?", while also providing a rebuttal interface through which the user can enter a grounded refutation whenever the AI's probe misses the mark, the success of a rebuttal here amounting to proof of epistemic agency. The Logic Status (Q4) visualises the state of the sixteen parameters in green (Resolved), yellow (Warning: ambiguous or inconsistent) and red (Null Error); the Stress Test Panel (Q4) predicts how scalability and reliability change under scenarios such as a hundredfold increase in data volume and probes accordingly; and the Compiled Story View (Q2) previews the chain of logic assembled from fragmentary Vibes and presents the limitations foreseeable before any experiment is run. The work in which the system warns that "Significance is Null" and the user extinguishes the error by putting beneficiaries and benefits into words is debugging of logic in the strict sense. The design philosophy of the UI is to guarantee, by compulsion, the logical rigour that contributes to the accumulation of scholarship, without slowing the speed of development that the Vibe carries. A concrete example of this screen design is shown in Figure 3: the Synthesis Pane corresponds to the Main Editor, the Analysis Sidebar listing the probes corresponds to the Sidebar, the Ontology Status Bar showing the fulfilment state of the parameters in green, yellow, and red corresponds to the Logic Status, and in addition the Agency Tracker, displaying the user's standing as a manager (Agency Status) together with a refutation score, makes the success of a rebuttal visible as proof of epistemic agency.

4.6. Outcomes: What the Compiler Demanded, and How the Authors Answered

This section organises the demands and probes that the Vibe Compiler actually issued in the process of building this paper's research logic, together with the authors' responses to them. This constitutes the only substantive evidence concerning the func-

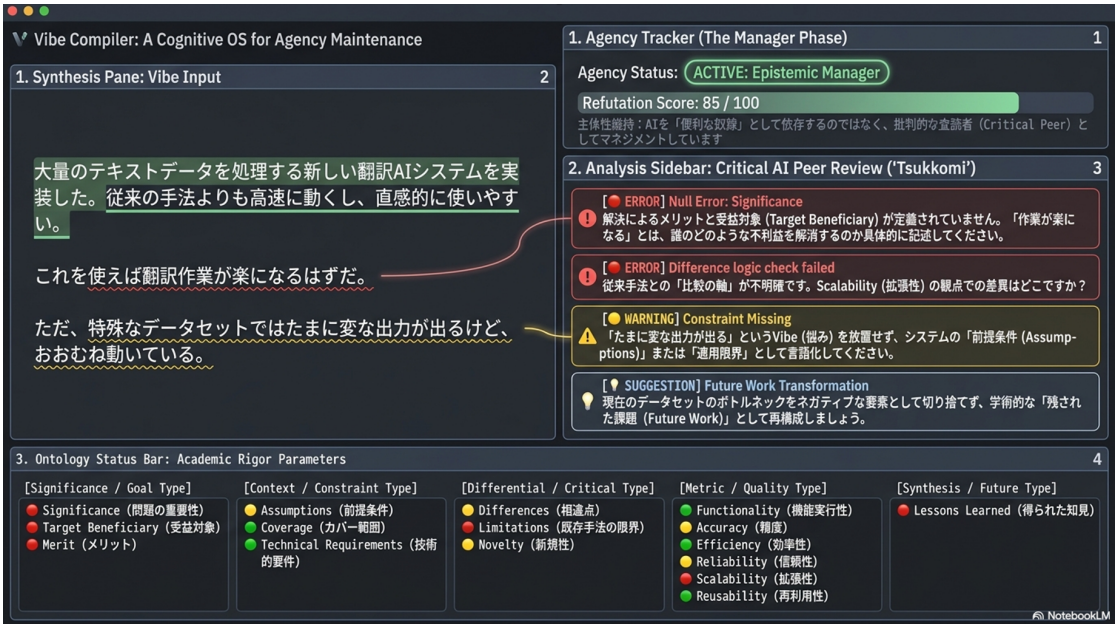

Figure 3: An example of the UI design of the Vibe Compiler (a scene in which a Vibe concerning research on a translation AI system has been entered). It comprises the Synthesis Pane into which Vibes are written, the Analysis Sidebar that lists the probes, the Ontology Status Bar showing the fulfilment state of the paper ontology, and the Agency Tracker showing the user's standing as a manager and a refutation score.

tional executability of the prototype, and at the same time a record that reciprocity at the second layer did in fact occur.

What is to be read out of Table 7 is that the compiler's demands were not mere requests to fill gaps. Of the ten, the naming of the concept (#3), the design of the evaluation indicators (#6) and the identification of the need for generalisation (#9) were points the authors had not prepared in advance and were produced only through reciprocity with the system. Item #9 is especially notable as a case in which the system pointed out the limits of its own scope of application, and in response this paper carried out the generalisation to three domain-general axes.

It is equally necessary to record what could have been lost in each demand. Once a Null has been named, the decision as to who fills that blank is the point at which the paths diverge. Had the user answered "write the description of the limitations for me," the settled logic would have become a product of the AI's Synthesis (Q2), and no judgement about what counts as sound (Q3) would have remained on the user's side. Even if the quality of the prose were much the same, who decided what to erect as the differential logic would have changed hands. What actually occurred was a sequence in which the user responded to the naming of a blank (Q4) with re-synthesis (Q1). What was protected is that decision, not the prose.

The functional executability of the prototype is thus confirmed. The system ran a type check against sixteen slots, detected unfilled slots as Null, returned feedback in the

Table 7: Principal demands from the Vibe Compiler and the authors' responses (a record of the process of building this paper's research logic)

#	Demand from the compiler (type-check error)	The authors' response and the logic settled
1	Why are existing cognitive models insufficient? Do they overlook S, or A, or the reciprocity?	Formulated the claim that existing models can describe the procedures and the products of problem posing but do not explain the dynamic process that deepens understanding. Settled the differential logic as "the black-box problem of the cognitive mechanism"
2	Who are the beneficiaries of this cognitive model? What new value do they obtain?	Settled that learners gain structural understanding and teachers gain the new, advanced expertise of supporting S&A reciprocity
3	If a distinctive name were given to this model, what image would it convey?	Coined "S&A reciprocity (Reciprocal Synthesis & Analysis)". Settled the definition that the mutual interchange is the source of the learning effect
4	Make the evaluative axes of Analysis concrete. What constitutes structural complexity?	Defined N_{step}, N_{var} and D_{map} for the arithmetic domain, and made the AI's solvability explicit as a premise at the same time
5	Is there no evaluative axis other than functional executability, the solvable/unsolvable distinction?	Added inquiry-driven Analysis, which evaluates positively the state of "being unable to solve it, yet being interested in the structure of its solution". A departure from the supremacy of the correct answer
6	How is the improvement of a learner's self-evaluation ability to be quantified?	Defined the four indicators E_{pred}, S_{cov}, L_{ref} and A_{epi} (Eqs. (3) and (4)). Proposed by the system itself and examined and settled by the authors
7	What is the decisive difference from existing problem-posing models?	Formulated the claim that, whereas Silver and Christou et al. grasped problem posing as an enumeration of procedures, this model focuses on the dynamic process of seeking to resolve dissonance
8	What is the difference from mere Reflection? As it stands this falls within an existing concept	Differentiated on three points: objective externalisation of the indicators, mutual delimitation of construction and evaluation, and deliberate maintenance of productive struggle. Settled "structural-gap-driven metacognitive support"
9	What are the implementation-level difficulties in extending S&A reciprocity to learners in general?	The system itself pointed out the difficulty of defining domain-general structural parameters. In response, generalisation to the three axes D_{depth}, W_{width} and M_{map} was carried out
10	The problem set out in the objective does not correspond logically to the insufficiency identified in existing methods	An instance of a consistency check firing. The objective of metacognitive support was brought into correspondence with the limitation of existing methods, their dependence on subjective noticing, and the solution vector of externalising objective indicators was synchronised with both

form of questions, and detected inconsistencies between slots. Here lies the affirmative evidence for RQ1. That the same mechanism operated in both the learner domain and the researcher domain with nothing changed but the Analysis mapping constitutes affirmative evidence for RQ4. And this paper, as an artefact, is itself a demonstration that a researcher's Vibe can be compiled into substantive research while a paper is generated from it. RQ2, RQ3 and RQ5, which ask about the magnitude of the effect, require measurement under controlled conditions and are therefore organised, as a matter of scope, in Section 6.5.

5. Application example: extending the model to the first layer (learners)

The execution log presented in Section 4 belonged to the second layer, that of the researcher. This section turns to the first layer, that of the learner, and what demands attention here is that the origin of the gap in the learner layer differs from the origin in the layer of research activity (Figure 4). Where the layer of research activity took the AI's Synthesis as the origin, what this paper designs for the learner layer is a configuration in which an external Analysis measures the human's Synthesis. The learner never converses with the generative AI, and the AI withdraws into the background as the machinery that computes the indices, so that the Synthesis of the support system is deliberately left unconnected to the learner. This is not an avoidance of generative AI but a design choice already justified in the layer of research activity, and the judgement it embodies is that of refusing to hand the learner a configuration of wholesale delegation.

The learner-layer sequences presented below are simulations derived from the design of the mechanism this paper describes, not records obtained from actual learners, and the values that appear in them are the values the design anticipates. What is at issue is not whether an artefact was produced but what can be lost in the course of producing it and what remains on the human side.

This configuration is placed at the same coordinates as Type (III) in Table 2, since Synthesis is carried out by the learner alone while Analysis is carried out by the learner's own self-assessment and by the support system's computation of indices side by side. Its origin differs from that of Type (IV), which this paper adopted for the layer of research activity, and in that sense the two layers do not share the same type. That the learner layer nevertheless does not remain within the limits of Type (III) follows from a condition added by design, namely that the learner declares the hunch about difficulty outwardly at the same moment as posing the problem, and that the judgement of which side the resulting difference is to be attributed to is left with the learner. Only once the hunch has been externalised does the difference from the index become an observable quantity, and the act of adjudicating where that discrepancy comes from, of deciding whether the way of seeing was too lenient or the problem that was posed was shallow, opens a re-interrogation of the very thing the learner set out to make (Section 5.3). Type (III) stops at the correctness of the assessment because it lacks a path that returns this adjudication of attribution to the user, and so the reach differs even where the coordinates coincide.

The difference from conventional support for learning by problem posing appears at exactly this point. Diagnosis by the single-problem structure model (Hirashima et al.,

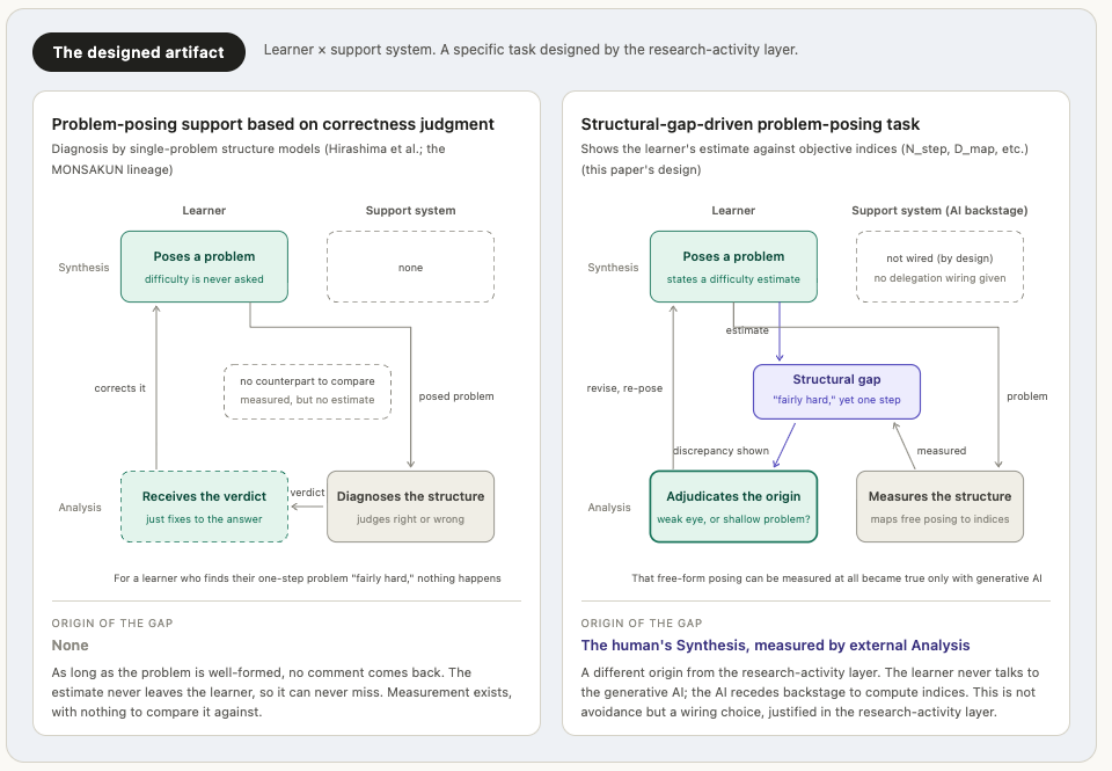

Figure 4: Contrasting configurations in the learner layer. In conventional support for learning by problem posing based on correctness judgement (left), the hunch never leaves the learner, so there is no origin for a gap. In the structural-gap-driven problem-posing task designed in this paper (right), the learner is made to state the hunch first, and an external Analysis measures it.

2008; 2014) judges whether the structure of the posed problem is correct, but when the problem is correct as a problem, no remark is issued. Because the learner's hunch never leaves the learner, it can never turn out to be off the mark. Measurement is present, yet there is nothing for it to be compared against, and for a learner who takes a self-made one-step problem to be "moderately difficult," nothing happens at all. The task designed here, by contrast, has the hunch declared in advance and then presents the difference from the objective index. That free-form problem posing became measurable at all only through generative AI is, moreover, the reason this design is specific to the age of generative AI.

5.1. The arithmetic problem-posing domain

The sequence begins with a learner who poses the problem "I bought an apple costing 100 yen. When a consumption tax of 10% is charged, how much is the payment?" (v1.0) and assesses it as "moderately difficult, because it requires a calculation." The system first confirms that a unique correct answer can be derived, which is solvability for the AI. Nothing so far departs from ordinary answer-oriented support; what follows does.

Scanning the structure of the posed problem, the system computes $N_{step} = 1$ (the single operation 100×1.1), $N_{var} = 1$ (the payment alone), and a low D_{map} (the expression can be built in the order given), and it thereby detects a structural gap against the learner's subjective declaration.

The system never offers the revision "add a discount." It throws the gap back in the form of a question instead: "The AI has derived the correct answer (110 yen) for the problem you posed. Against the goal of structural complexity you yourself set, however, the current number of operational steps is one. How would this change if you increased the number of unknowns, or built into the context an operational component running in the opposite direction, such as a discount?" To give only the gap rather than the answer, or to give the revision itself, is the point at which the twin demands of protecting and driving take shape as a concrete design decision, and the moment the latter is chosen, productive struggle is erased (Hiebert & Grouws, 2007). The learner then sets the goal of deliberately raising the difficulty and reconfigures the components (Table 8).

Table 8: Transition of the structural parameters before and after one cycle of reciprocity in the arithmetic problem-posing domain (simulation)

Version	Gist of the problem statement	N_{step}	N_{var}	D_{map}	Learner's self-assessment
v1.0	One apple at 100 yen with 10% consumption tax	1	1	Low	"Moderately difficult, because it requires a calculation"
v2.0	Three apples at 100 yen each, a 20% discount, 10% consumption tax, and change from a 500 yen coin	4	2	Medium to high	"It became harder because I added the change"

Note: as stated at the beginning of this section, the values in this table are a simulation based on the design.

What matters is that the reciprocity does not end at v2.0. Once the learner has put the insight into words, saying "Adding the change means that I have to work out not only the payment but also a subtraction step, and above all that I have to consider the relation between the payment and the money I hold, so the mapping becomes harder. Is this what it means for D_{map} to rise?", the system records this as a logic snapshot and returns further probes. The probe that presses on the fragility of the premises asks, "This problem depends on the premise that the 500 yen coin always suffices. Does the expression still function as it stands if the unit price is rewritten as 200 yen?" The probe that addresses the ambiguity of the order of application asks, "Which is to be applied first, the 20% discount or the 10% consumption tax? As the statement currently reads, this is not uniquely determined." The probe that addresses extensibility asks, "If the apples became 100 in number and a bag charge and a points rebate were added, how would the structural complexity change?"

What this two-stage arrangement shows is that probing a v2.0 with which the learner is already satisfied exposes issues that could not even have been raised at the stage of v1.0. The mutual delimitation whereby the output of Analysis becomes an argument to

the next Synthesis is reiterated in precisely this form. None of the three probes stops at naming a quality index; each translates the index into components and numbers of the artefact, the unit price, the order of application, and the number of items. What separates a probe that functions as a critical file, in the sense of the rasping tool, from one that does not is not its topic but its granularity, and the moment the same issue is stated at the level of "make the premises explicit," the reciprocity spins idle.

What stands to be lost here is that the learner may finish without ever noticing that the self-assessment "moderately difficult" departs from the actual structure. The posing of the problem itself succeeds, which is to say that Q1 succeeds. The self-assessment nonetheless fails to agree with the structure, which is to say that Q3 fails. Under a description that lacks the double distinction, this state can be written down only as "a reasonably good problem was produced." What was protected is the goal setting of redefining difficulty for oneself, not the value of N_{step} as such. Should the learner reply "then fix it for me," v3.0 will duly come into being, but the recognition of what the learner had overlooked will not remain.

5.2. The Japanese reading comprehension domain: only the indices are replaced, the mechanism is not

In the Japanese language domain the components of Synthesis become keywords, connectives, paragraph blocks, and contrastive structures, and the structural indices are replaced by S_{ref}, L_{link}, and V_{map}. Only the symbols are exchanged while the mechanism remains as it was, and it is precisely the ease of this replacement that underwrites the domain generality of the model. Taking as the source text an expository passage arguing that AI is good at calculation but does not understand meaning, v1.0 reads "What is the weakness of AI? Extract it from the passage," and is computed as having a minimal S_{ref}, $L_{link} = 1$, and a V_{map} of zero. The self-assessment, however, is that this is a good problem because it asks about the basics, so a gap arises between the subjective sense and the structure. The system asks, "L_{link} is at its minimum of one. How would L_{link} change if you combined the question with the human-specific embodiment discussed in the adjacent paragraphs and reshaped it into a form that requires the reason to be explained?" The reconstructed v2.0 reads "Drawing on the claim made in the passage, explain in no more than 40 characters why AI can be said not to understand meaning, comparing it with the characteristics of human embodiment," and is computed as having a large S_{ref}, $L_{link} = 3$, and a high V_{map}. What was exchanged between arithmetic and Japanese is only the way each axis is measured, while the form of the reciprocity and the requirement on the granularity of the probes are identical.

5.3. Reverse Analysis: how the AI's mis-assessment turns into a learning opportunity

Reading the AI's errors not merely as defects but also as occasions for observing where agency resides is characteristic of the model. Reverse Analysis works head-on, however, in the layer of research activity, where the user converses with the AI directly, which is Type (IV). When a researcher objects with grounds that "the limitation of the existing method the AI computed overlooks a domain-specific constraint," the system

accepts this as reverse Analysis and records it as a high-quality objection that identifies a contradiction among the premises and offers an alternative criterion. The A_{epi} of Eq. (4) counts objections of this kind, weighting them by their quality.

The counterpart in the learner layer is not an objection to the AI but the act of adjudicating where the discrepancy comes from. A learner confronted with the difference from the objective index has to determine whether the way of seeing was too lenient or the problem that was posed was shallow. Determining the former leads to remaking the hunch, and determining the latter to remaking the problem. The judgement of which side the discrepancy is attributed to is left with the learner, and this is what Q3 amounts to in the learner layer. If the learner can ground the lowness of the index in a design intention of the learner's own, saying that the readers are elementary school pupils and that keeping the wording deliberately plain was given priority, that constitutes a judgement attributing the discrepancy not to the shallowness of the problem but to a difference in the premises.

So long as the computation of the indices is entrusted to a language model, erroneous measurement is unavoidable in principle (Huang et al., 2025; Ji et al., 2023). The configuration that keeps the AI in the background in the learner layer also has the effect of narrowing the paths by which such errors flow directly into the learner's judgement.

That an erroneous assessment is returned is unavoidable in principle so long as the computation of the parameters is entrusted to a language model (Huang et al., 2025; Ji et al., 2023). Concealing this and letting the system behave as an oracle leads users to abandon their own correct intuitions and follow the AI, and given the tendency of large language models to agree excessively with their users (Sharma et al., 2023), this danger is a realistic one. The model therefore positions the AI as an imperfect partner that errs at times yet stimulates thought, and it makes keeping the path of objection permanently open a design requirement. This is a demythologisation that accords with the demands of critical digital pedagogy (Stommel et al., 2020). What is protected here is not the artefact, whether a summary or a posed problem, but the locus of the judgement of what counts as valid.

5.4. Positioning in the four quadrants, and an explanation of learning by problem posing

The cases above can now be placed in the four quadrants (Table 3). In the second-layer session recorded in the execution log, the injection of intuition and the resynthesis fall in Q1, the naming of Nulls by the type check falls in Q4, and the judgement of what should be written falls in Q3, forming the sequence Q1 $\rightarrow$ Q4 $\rightarrow$ Q3 $\rightarrow$ Q1. The arithmetic and Japanese domains take the same sequence, with the posing of the problem in Q1, the self-assessment in Q3, the computation of the structural parameters and the probing in Q4, and the reconfiguration in Q1. The objection to a mis-assessment is a case in which Q3 has countered an erroneous Q4. In bare use of generative AI, the capable-servant mode, the user's Q1 degenerates by contrast into the composition of a request, Q2 takes over the selection and combination of components, and Q3, which assesses the validity of the artefact, is left empty.

To see what becomes indescribable if this distinction is not introduced, the second-layer episode may be written down in two ways. One description reads that the type

check named three Nulls (Q4), that the user judged what to take as the differentiating logic (Q3), and that a settled logic was produced (Q1). The other reads that the author completed the introduction with the support of AI, the division of labour being that AI generated the prose while the human decided the direction. Under the vocabulary the latter employs, that of division of labour, task allocation, cognitive offloading, and attribution of responsibility, whether Q3 was executed never appears in the description, and the process collapses into the single event of an introduction having been written with AI support. For the same reason the capable-servant mode and the sequence described here look identical in that an artefact was produced using AI, whereas the four quadrants describe them separately, the one as a sequence in which Q1 degenerates into the composition of a request, Q2 takes over the combination, and Q3 is left empty, the other as a sequence in which Q1, Q3, and Q4 are all present. The state in the arithmetic domain in which the posing of the problem itself succeeds while the self-assessment fails to agree with the structure can likewise be separated out only as a state in which Q1 succeeds and Q3 fails. What distinguishes the human/AI axis from theories of the division of labour is that it is placed there not in order to optimise allocation but in order to name the deficiency that is the blank at Q3. The absence of this vocabulary is more than a descriptive inconvenience, for what cannot be named as being lost can be neither designed for nor assessed.

The positioning set out above offers an explanation of the mechanism by which learning by problem posing takes effect. That problem posing promotes learning has been reported repeatedly (Cai et al., 2015; Silver, 1994), yet why it does so has not been adequately explained (Cai & Hwang, 2020). Seen through the four quadrants, problem posing differs from ordinary problem solving in that it imposes on the learner both Q1, the decision of what to ask together with the combination of components, and Q3, the assessment of the learner's own artefact, at one and the same time. Externalising Q4 in addition produces a divergence between the results of Q3 and Q4, and that divergence becomes an argument to the next Q1. Problem posing promotes learning, in other words, because it is one of the few activities that impose Q1 and Q3 on the same learner simultaneously. This explanation responds at the level of mechanism to Silver (Silver, 1994), who positioned problem posing as an activity before, during, and after problem solving while leaving the cognitive process of problem posing itself at an abstract level, and to Christou et al. (Christou et al., 2005), who described it as an enumeration of procedures. The explanation moreover yields testable predictions. Conditions that separate Q1 from Q3, namely a condition in which learners pose problems but do not assess them and a condition that imposes the self-assessment alone, should show a diminished learning effect, and whether Q4 is externalised should govern whether learners can become aware of their own divergence. The model thus supplies research on learning by problem posing with a design guide for such contrastive conditions.

6. Discussion

6.1. Content-orientation reinstated: an unformalised ontology ran on an LLM

Of everything this paper yields, the finding with the widest reach is the implication of the fact reported in Section 4 (Section 4.2): what drove the Vibe Compiler was not the inference engine but the structure of the content we supplied to it.

When Bourdeau and Mizoguchi (2000) argued that every difficulty obstructing the construction of intelligent educational systems is a matter of content, and that neither inference technology nor elegant theoretical formalisation improves the situation, the premise tacitly at work was that content must pass through a stage of formalisation before a computer can handle it. A quarter of a century of ontological engineering may fairly be described as the pursuit of how to carry out that stage robustly. When Bourdeau and Mizoguchi (2016) restated that the distinctions an ontology draws are not a question of how they are represented on a computer, the remark warned against an understanding that tends to collapse ontology into techniques of formalisation. What our prototype shows is that the warning was right in a way it had not anticipated. The paper ontology we supplied was not a description in a formal language but a prose document written for human readers, and constructing a formal ontology out of it would ordinarily have been a prerequisite. With the arrival of generative AI, however, it ran simply by being fed into NotebookLM. The quality of its operation, moreover, did not stop at the shallow level of testing whether items exist; it reached the level of actually rejecting failing cases that cannot be detected without examining the relations among items.

The propositions that follow from this connect in stages. The point to affirm at the outset is that the value of heavy-weight ontologies for building robust and verifiable systems has not diminished in the slightest: for guaranteeing consistency, the soundness of inference, and reuse, formalisation remains irreplaceable. Yet any content-oriented structure carrying constraints can now be reasoned over by an LLM without passing through formalisation, and the paper ontology is precisely such a case. The consequence is single. The centre of value has shifted from the skill of formalising structure toward the side of content, toward the question of what ought to be given structure at all.

This shift does not compete with ontology research; it raises that research's claim again under different conditions. The importance of content-orientation was once argued in order to justify the cost of formalisation. Now that part of that cost has fallen away, the quality of the content itself directly determines the outcome. We restate this turn from the standpoint of protecting agency, because the question of what to give generative AI is inseparable from the question of what of the human it replaces and what it excites. Where no structure is given, fluent prose comes out and no gap arises, and where no gap arises metacognition is not driven. Designing content in the age of generative AI is thereby the design of human agency itself. The practical consequence is plain: what is needed to drive generative AI as intended is not the artistry of prompt engineering but the design of the content supplied to it. This claim descends to a testable prediction. It is for this reason that the reproduction procedure of this paper is given not as a manual of operations but as a list of the materials supplied (Section 4, 4.1): feeding the same set into a different reasoning environment should reproduce the same type-checking

behaviour. Isolating how far features specific to NotebookLM contribute remains a task for future work involving experiments.

6.2. Using generative AI by "educating" it

The implications of content-orientation have a further side. What actually took place on the way to the Vibe Compiler was an act of persuading the generative AI. As the dialogue history collected in the appendix shows, the AI initially maintained that "in supporting paper writing and the conduct of research, nothing corresponding to code in Vibe Coding can be generated," and it grounded this in the observations that an academic paper is not a mere artefact of expression but the presentation of a delta against an existing body of knowledge together with a logical warrant, that code possesses an immediate feedback mechanism whereas research logic possesses none, and that autonomously completing the antinomic process of synthesising a new claim while criticising it at the same time is difficult. The authors answered by supplying what was missing. We gave the AI the "type" constituted by the paper ontology, the input channel from the human side that we call Vibe, and the 16 parameters as axes of evaluation. The AI then revised its conclusion, holding that "because the AI already knows what counts as correct (the type of a paper), pouring in what the human wants to do (the Vibe) gives it ample potential to function as a compiler that automatically synthesises the research logic bridging the two."

This episode contains an element that generalises as methodology. The limits of a generative AI's capability often appear not as intrinsic limits of the model but as a function of premises that have not been supplied. Asking a generative AI that declares it cannot do something what it lacks, and then supplying the missing element as content, is a technique of a lineage quite distinct from prompt engineering. We position this as a way of using generative AI that educates it. Here again what does the work is not the inference engine but the content, so the claim of the preceding section finds support in this episode as well.

6.3. Theoretical, methodological, and practical implications

Our model contributes concretely to several theoretical lineages. To research on learning by problem posing it supplies an account of the long-unexplained mechanism by which posing problems promotes learning (Cai & Hwang, 2020), namely that the activity imposes Q1 and Q3 on one and the same learner simultaneously. To research on reflection it structures the vague notion of introspection, which has depended on subjective noticing (Schön, 1983), as a dissonance between Synthesis and Analysis, and redefines it as a dynamic mechanism accompanied by collation with objective indicators and immediate reconstruction; positioning evaluative judgement (Tai et al., 2018) not as a capacity for evaluation but as an engine that drives the next action is a further contribution to this lineage. To research on epistemic agency it extends a concept that has been discussed as the sharing of responsibility among humans (Damşa et al., 2010; Scardamalia & Bereiter, 2014) into the space between humans and AI, and operationalises it as an observable behaviour, a rebuttal to the AI's evaluation for which grounds have been

established, thereby giving a measure to what had remained abstract. The trichotomy of Cox (2024) is likewise brought down from a normative framework to the level of a theory of mechanism, where the question becomes what kind of dialogue design brings about the transition from Makers to Managers. To the design theory of hybrid intelligence, finally, it introduces a design variable that does not reduce to the maximisation of outcomes: the perspective that asks how the four powers are redistributed to the human side (Akata et al., 2020) takes concrete form as the design guideline of withholding Q3.

At the methodological level, the procedure of treating a paper ontology as a type system applies beyond the confines of this paper, since it renders the pointing-out of what is missing, done tacitly in research supervision and in peer review, as the explicit procedures of checking slot fulfilment and checking consistency among slots. The consistency checks in particular (Table 6) differ from checklist-style guidance in that they reject documents that a test for the mere existence of items would pass. Our four indicators likewise serve not only for evaluating learners but, for researchers developing systems, as debugging indicators for support logic. The stance of rereading an AI's mistaken evaluation as an occasion for observing agency rather than as a defect is also transferable. In the evaluation of generative-AI-based support systems, errors are ordinarily treated as noise to be reduced; we have instead woven them into the design and converted them into the requirement that the path of rebuttal stay open at all times.

At the level of practice, our model becomes an instrument for converting a supervisor's tacit knowledge into explicit procedure. Questions such as what the significance of a piece of research is, or what is insufficient about existing methods, have conventionally been posed out of the supervisor's experience. The 16 parameters and the consistency checks externalise the system of these questions, and one conceivable practice is for a student to pass through the Vibe Compiler before a meeting and finish putting the Null slots into words. In designing lessons around learning by problem posing, presenting structural parameters makes concrete an instruction that could previously say no more than "pose a difficult problem," and it asks of the teacher a shift from the role of delivering correct answers to the role of gauging the divergence between the learner's self-assessment and the actual structure and designing probes at an appropriate grain size, which is to say a new expertise consisting in the support of S&A reciprocity. In drafting policies on AI use, the four quadrants replace the dichotomy of prohibiting AI or allowing it freely with provisions written quadrant by quadrant, of the form "Q2 is permitted, and Q3 must be carried out by the person concerned and recorded."

6.4. Critical examination

The perspective of critical digital pedagogy (Morris & Stommel, 2018; Stommel et al., 2020) has pointed out that support technologies aimed at efficiency can restrict learners' agency and reproduce existing inequalities. Our position, which actively incorporates generative AI into educational settings, is not exempt from this critique. Our model operates at the level of structuring thought and does not itself dissolve inequality of access to high-performance generative AI models; there are reports that generative AI can instead widen the gap among novices (Prather et al., 2024), so learners who struggle

with metacognition may be the least likely to benefit. Reliance on the average and biased knowledge that AI generates can moreover induce a monoculture of scientific knowledge (Messeri & Crockett, 2024), and although we attempt to resist this by placing the individual's intuition at the origin of the logic, we are not free of that bias so long as parameter computation and probe generation are entrusted to a foundation model. The very criterion by which something is judged "structurally shallow" may reflect the model's skew.

A more fundamental issue is that the boundary separating the externalisation of thought from its replacement is not sufficiently defined, either theoretically or empirically. As research on cognitive offloading (Risko & Gilbert, 2016) shows, externalisation can both release cognitive resources and atrophy capability. The distinction between accelerative and exploratory modes of use (Barke et al., 2023) lies close to our view, but we have yet to answer the objection that the repetition of reciprocity may itself become a new form of dependence. If mistaken parameter values or counterfactual limitations of existing methods are presented, furthermore, epistemic trust is damaged and the will to inquire declines, and no measured value for the rate of mistaken evaluation exists. Intervention at an inappropriate moment obstructs the user's authority to decide and can produce the paradox of support that violates the very agency it is meant to protect. Even in the execution log of the second layer we observed that when type checking fires exhaustively the resulting remarks can break the continuity of the work. So long as protection and driving are demanded at once, comprehensiveness of support and continuity of work stand in a trade-off and firing must be controlled according to stage, yet that control law does not follow from our model.

6.5. Limitations

We begin with the self-referential character peculiar to this paper. Much of what is written here derives from the output of a prototype built on the very mechanism this paper describes. What can be stated explicitly as human judgement is the framing of the problem, the selection of what to discuss, the rebuttal and rejection of the system's output, and the decision on the overall structure. Much of the drafting, the choice of words, and the development of examples is the system's output, and the authors occupied the position of judging whether to adopt it. This division of labour is precisely the reciprocity of the researcher layer as we describe it, and this paper is a case of the self-application of its own claim. That fact is what constitutes the demonstration that Vibe Compiling is possible. What self-application shows, on the other hand, is that the mechanism actually runs, not that it produces a learning effect, and the separation of the two is a boundary we have drawn deliberately. To the question of whether this is wholesale delegation in substance while claiming to protect metacognition, Table 7 shows in traceable form which piece of logic was produced as a response to which type-check error, and we invite readers to examine the matter against that record. Controlling for the confirmation bias that arises when evaluator and evaluated are the same party is a task for the next stage.

We list the remaining limitations below.

- A single case, and one in which the authors are themselves the users. The record of the researcher layer rests on the single case of the process of building this paper's research logic, and no device for excluding confirmation bias and self-assessment bias is built into the design.
- No quantitative data for the learner layer exist. No controlled experiment has been conducted, and the changes in E_{pred}, S_{cov}, L_{ref}, and A_{epi} together with transfer effects are all unverified predictions. We make no quantitative claim about the size, the conditions, or the persistence of any effect.
- The prototype's configuration depends on off-the-shelf services. Being a combination of commercial services, its behaviour can change with changes in their specifications. Because parameter computation depends on generation by a large language model, it is not deterministic, and identity of output for identical input is not guaranteed.
- The examination is confined to a single language, Japanese. Indicators such as the vocabulary substitution difficulty V_{map} depend strongly on linguistic structure, and their viability in other languages is unconfirmed.
- Domain-specific parameters are defined by hand. Designing the Analysis mapping requires the judgement of a domain expert. The generality at issue is a generality of the mechanism rather than a generality of the settings, and a procedure for validating the parameters is not yet in place.
- Structural parameters do not guarantee the semantic soundness of an artefact. High values of N_{step} or L_{link} do not entail that the artefact holds up as a problem. A shopping word problem in which the unit price is changed from 100 to 200 yen admits the exception that the amount cannot be paid with a 500-yen coin, and it ceases to hold without an added precondition. The danger that probes issued from the side of structure overlook a breakdown on the side of meaning remains.
- The fragility of the premise that the AI can solve the task. This premise breaks the more readily as difficulty rises, and once it breaks the objective value that serves as the reference for E_{pred} is unavailable. Should the fact of the break be concealed by hallucination (Huang et al., 2025; Ji et al., 2023), a self-assessment may be judged to be off the mark against an erroneous reference value.
- The transfer of evaluative judgement remains a hypothesis. The implication that a stance attentive to structure is carried over across domains is a prediction from our model, and no data supporting transfer exist.

One further point deserves mention. That among the materials we fed into NotebookLM was a survey paper on agency (Cox, 2024) conditions the context in which this paper applies. Feeding in a different survey paper would yield a different context for the use of generative AI, and the system of probes the compiler issues would change

accordingly. This does not narrow the range of the paper's value, however, because the degradation of agency that follows from using generative AI as a capable servant is not the claim of one particular survey paper but a widely shared and general problem. The property that the system adapts to any context through replacement of the supplied materials is itself a corollary of the content-driven finding we advance.

7. Conclusion

7.1. Summary

The rise of generative AI threatens to reduce the construction and transmission of knowledge to the efficient processing of information, and it puts in question whether human beings can continue to secure epistemic agency. The problem this paper has addressed is how to design and realize a support mechanism that, in the course of converting an inchoate intuition (Vibe) into academic logical rigour, preserves the human's authority to decide while honing metacognition. Against this problem the paper has proposed structural-gap-driven metacognitive support, which treats the dissonance between the user's subjective construction (Synthesis) and the objective structural indices the system presents (Analysis) not as an error to be removed but as a source of metacognitive stimulation. This idea is carried by the S&A reciprocity model, which construes intellectual construction as a mutually constraining reciprocity between Synthesis and Analysis; by the dual-layer structure that unfolds the model across the layer of the learner and the layer of the researchers themselves; and by a design that configures AI not as an entity that supplies answers but as a critical file, in the sense of the rasping tool, that bears down on the fragility of premises. The prototype Vibe Compiler is the implementation of all of these.

The findings we would ask the reader to take away are condensed into the following five points.

1. Generative AI makes it possible to compile research logic and to write a paper semi-automatically. This paper is itself a product of that process. Vibe Compiling, as distinct from Vibe Coding, is already feasible work.
2. That capability can contribute to the enhancement of the metacognitive function by which the crisis of agency is met. So long as type-check errors are returned in the form of questions rather than answers, the user retains productive struggle and follows a path that raises them from creator to manager.
3. A self-application takes shape in which the crisis brought about by generative AI is met by a way of using generative AI. The model proposed here is applicable to itself, and the second layer is precisely that application in execution.
4. Without passing through prompt engineering, feeding solid content into NotebookLM is enough to steer generative AI this easily. This is the principal take-home lesson of the paper. Ontological documents written for human readers, never formalized, ran on the LLM as the very skeleton of its inference. The centre of value has shifted from the skill of formalizing structure toward the content side, toward the question of what ought to be given structure.

5. One and the same mechanism operates in the learner domain and in the researcher domain alike, with nothing replaced but the Analysis mapping. Its application to three heterogeneous domains, problem posing in arithmetic, reading comprehension, and research-logic synthesis, supports this generality.

The four types of structural gap function above all as a vocabulary for designing and evaluating support systems in the age of generative AI, because they make it possible to ask of any system, "By what function, and against whose Synthesis or Analysis, does this system attempt to excite a structural gap?" The era in which humans criticized their own work unaided, the situation in which wholesale delegation to AI makes the gap vanish, the framework in which humans and AI generate a gap out of an Analysis they execute together, and the route taken here, in which the AI's Analysis probes the results of the AI's Synthesis and excites the human's metacognition: only by distinguishing these four does the design intent of each system become describable.

7.2. Problems to Be Solved

The problems set out below are the work that must next be carried out in order to fix the scope of the model, and they form three stages. The first stage asks what is to be carved out as the object of Synthesis in the first place; the second asks which indices can be extracted from that class so that the activity in question becomes observable; and the third asks how those indices are to be operated and controlled if a valid and reproducible interaction is to be realized. What the third stage demands is not precision of measurement. Even where measurement is somewhat coarse, the question is whether an interaction holds that is valid in the sense that metacognition remains on the human side, and reproducible in the sense that others can replicate it.

Stage 1: Delimiting the Class of Applicable Activities. The four quadrants need to be applied beyond the small number of cases treated here, to the diverse situations in which generative AI is used, and it must be organized across cases which cell has to be reserved for the human if epistemic agency is to be preserved. This work accumulates explanatory power for the conceptual distinction independently of any verification of effects, and it draws out inductively the contours of the class of activities that can be described as S&A reciprocity. The range of application should at the same time be extended to component-assembly work such as programming, experimental design, and design, so as to test whether S&A reciprocity holds as a general model. Code generation by generative AI (Karpathy, 2025; Sarkar & Drosos, 2025) is the domain in which the problem taken up here appears in its sharpest form.

Stage 2: Deriving and Validating the Indices. The objective parameters computed by the Analysis mapping are at present defined by hand for each domain, so the generality this paper can claim is a generality of the mechanism rather than a generality of the settings. If axes of structural complexity could be derived semi-automatically from a set of artefacts, and the generation of the lower layer automated while the two-layer structure is preserved, the claim of generality would be raised from the level of design to the level

of implementation (this task is one that the prototype itself pointed out as its own limitation; #9 in Table 7). A mechanism is also required by which the AI discloses from which components and in what manner it counted the structural parameters, declares of its own accord the range within which it can evaluate correctly, and issues a warning when that range is exceeded. This self-declaration is a requirement for not concealing the moments at which the Oracle premise breaks down. Beyond this, the bias that the cultural and linguistic bias of the foundation model imparts to the computation of the structural parameters must be examined comparatively across several models and several languages, for if criticism issues from a single averaged standpoint, the file can become an instrument that grinds away the diversity of intuitions.

Stage 3: Verifying Effects and Establishing Reproducibility. Whether the presentation of a structural gap hones learners' evaluative judgement and sustains productive struggle can be verified only by comparison under controlled conditions. Measuring the change in the four indices under such conditions is the task of highest priority, and the transfer hypothesis must be examined alongside it, through a post-task asking whether learners who have experienced reciprocity in arithmetic problem posing increase their references to structure when posing problems in reading comprehension. The effect of repeated reciprocity on abstract thinking ability, and the question of whether evaluative ability is internalized and persists once the support is withdrawn, require a longitudinal design that includes a delayed post-test. Waiting here is the sharpest objection of all, namely that the repetition of reciprocity may itself become a new form of dependence, and comparison with existing approaches that cultivate generative-AI literacy through self-regulated learning (Anders & Dux Speltz, 2025) belongs to this context as well. It is equally indispensable to examine, as design research, how a teacher can operate S&A reciprocity in ecologically valid settings such as an actual classroom or laboratory (Cai & Hwang, 2020). The dependence on off-the-shelf services must also be shed: an implementation that externalizes the ontology definitions, the conditions for consistency checks, and the management of logic snapshots has to be released, so that versions can be fixed and third parties can replicate the work. Since the design involves the deliberate presentation of erroneous evaluations, moreover, ethical review and debriefing are inseparable requirements.

7.3. Closing Remarks

Generative AI can be configured as a device that substitutes for human intellectual activity, and equally as a device that excites it. What decides which it becomes is not the capability of the model but what we give to it. Vibe Compiler was able to keep returning type-check errors because it had been given content in the form of a paper ontology, and because that content was not a list of questions but a structure carrying conditions between one question and another.

The final claim of this paper is therefore the following. In the age of generative AI, the work of protecting human epistemic agency lies neither in keeping AI at a distance nor in making AI cleverer, but in designing what is given to AI. The dual-layer model of

S&A reciprocity is the framework that guides that design, and Vibe Compiler is its first implementation. This paper itself was written by means of that framework.

Disclosure statement

The authors declare no competing interests.

CRediT author statement

Riichiro Mizoguchi: Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Project administration. **Tomoki Aburatani**: Conceptualization, Formal analysis, Writing - review & editing, Visualization. **Kento Koike**: Conceptualization, Investigation, Writing - review & editing, Visualization. **Machi Shimmei**: Investigation, Writing - review & editing.

References

Ajjawi, R., Tai, J., Dawson, P., & Boud, D. (2018). Conceptualising evaluative judgement for sustainable assessment in higher education. In D. Boud, R. Ajjawi, P. Dawson, & J. Tai (Eds.), *Developing evaluative judgement in higher education: Assessment for knowing and producing quality work* (pp. 7–17). Routledge.

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., ... Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, *53*(8), 18–28.

Anders, A. D., & Dux Speltz, E. (2025). Developing generative AI literacies through self-regulated learning: A human-centered approach. *Computers and Education: Artificial Intelligence*, *9*, 100482.

Asimow, M. (1962). *Introduction to design*. Prentice-Hall.

Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, *1*(2), 164–180.

Barke, S., James, M. B., & Polikarpova, N. (2023). Grounded Copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages*, *7*(OOPSLA1), 85–111.

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*, *49*(6), 893–905.

Boud, D., Keogh, R., & Walker, D. (Eds.). (1985). *Reflection: Turning experience into learning*. Kogan Page.

Bourdeau, J., & Mizoguchi, R. (2000). Collaborative ontological engineering of instructional design knowledge for an ITS authoring environment. *Intelligent Tutoring Systems (ITS 2000)*, *1839*, 399–409.

Bourdeau, J., & Mizoguchi, R. (2016). Ontological engineering for theory-aware educational systems. *International Journal of Artificial Intelligence in Education*.

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences*, *2*(2), 141–178.

Cai, J., & Hwang, S. (2020). Learning to teach through mathematical problem posing: Theoretical considerations, methodology, and directions for future research. *International Journal of Educational Research*, *102*, 101391.

Cai, J., Hwang, S., Jiang, C., & Silber, S. (2015). Problem-posing research in mathematics education: Some answered and unanswered questions. In F. M. Singer, N. F. Ellerton, & J. Cai (Eds.), *Mathematical problem posing: From research to effective practice* (pp. 3–34). Springer.

Cai, J., Hwang, S., & Melville, M. (2023). Mathematical problem-posing research: Thirty years of advances building on the publication of "on mathematical problem posing". In J. Cai, G. J. Stylianides, & P. A. Kenney (Eds.), *Research studies on learning and teaching of mathematics* (pp. 1–25). Springer.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43.

Christou, C., Mousoulides, N., Pittalis, M., Pitta-Pantazi, D., & Sriraman, B. (2005). An empirical taxonomy of problem posing processes. *ZDM – Zentralblatt für Didaktik der Mathematik*, *37*(3), 149–158.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences*, *13*(1), 15–42.

Cox, G. M. (2024). Artificial intelligence and the aims of education: Makers, managers, or inforgs? *Studies in Philosophy and Education*, *43*(1), 15–30.

Cross, N. (2006). *Designerly ways of knowing*. Springer.

Cui, K. Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S., & Salz, T. (2024). The productivity effects of generative AI: Evidence from a field experiment with GitHub Copilot. *An MIT Exploration of Generative AI*.

Damşa, C. I., Kirschner, P. A., Andriessen, J. E. B., Erkens, G., & Sins, P. H. M. (2010). Shared epistemic agency: An empirical study of an emergent construct. *Journal of the Learning Sciences*, *19*(2), 143–186.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643.

Dorst, K., & Cross, N. (2001). Creativity in the design process: Co-evolution of problem–solution. *Design Studies*, *22*(5), 425–437.

Dubberly, H., Evenson, S., & Robinson, R. (2008). The analysis-synthesis bridge model. *interactions*, *15*(2), 57–61.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). Opinion paper: "so what if ChatGPT wrote

it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. MIT Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911.

Floridi, L. (2014). *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), 6.

Guilford, J. P. (1967). *The nature of human intelligence*. McGraw-Hill.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404, Vol. 1). Information Age Publishing.

Hirashima, T., Yamamoto, S., & Hayashi, Y. (2014). Triplet structure model of arithmetical word problems for learning by problem-posing. In S. Yamamoto (Ed.), *Human interface and the management of information: Information and knowledge in applications and services* (pp. 42–50, Vol. 8522). Springer.

Hirashima, T., Yokoyama, T., Okamoto, M., & Takeuchi, A. (2008). An experimental use of learning environment for problem-posing as sentence-integration in arithmetical word problems. *Intelligent Tutoring Systems (ITS 2008)*, *5091*, 687–689.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.

Karpathy, A. (2025, February). *There's a new kind of coding I call "vibe coding"*. X (formerly Twitter). Retrieved July 30, 2026, from https://x.com/karpathy/status/1886192184808149383

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task.

Lee, H.-P. (, Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *Pro-*

ceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), 1–22.

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58.

Mizoguchi, R. (2004). Tutorial on ontological engineering—part 3: Advanced course of ontological engineering. *New Generation Computing*, *22*(2), 198–220.

Morris, S. M., & Stommel, J. (2018). *An urgency of teachers: The work of critical digital pedagogy*. Hybrid Pedagogy Inc. https://criticaldigitalpedagogy.pressbooks.com/

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422.

Perry, N., Srivastava, M., Kumar, D., & Boneh, D. (2023). Do users write more insecure code with AI assistants? *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, 2785–2799.

Prather, J., Reeves, B. N., Leinonen, J., MacNeil, S., Randrianasolo, A. S., Becker, B. A., Kimmel, B., Wright, J., & Briggs, B. (2024). The widening gap: The benefits and harms of generative AI for novice programmers. *Proceedings of the 2024 ACM Conference on International Computing Education Research (ICER '24), Volume 1*, 469–486.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688.

Sarkar, A., & Drosos, I. (2025). Vibe coding: Programming through conversation with artificial intelligence.

Scardamalia, M., & Bereiter, C. (2014). Knowledge building and knowledge creation: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), *The cambridge handbook of the learning sciences* (2nd ed., pp. 397–417). Cambridge University Press.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models [Published at ICLR 2024].

Silver, E. A. (1994). On mathematical problem posing. *For the Learning of Mathematics*, *14*(1), 19–28. https://www.jstor.org/stable/40248099

Sowden, P. T., Pringle, A., & Gabora, L. (2015). The shifting sands of creative thinking: Connections to dual-process theory. *Thinking & Reasoning*, *21*(1), 40–60.

Stommel, J., Friend, C., & Morris, S. M. (Eds.). (2020). *Critical digital pedagogy: A collection*. Hybrid Pedagogy Inc.

Stroupe, D. (2014). Examining classroom science practice communities: How teachers and students negotiate epistemic agency and learn science-as-practice. *Science Education*, *98*(3), 487–516.

Supianto, A. A., Hayashi, Y., & Hirashima, T. (2017). Model-based analysis of thinking in problem posing as sentence integration focused on violation of the constraints. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1–21.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, *76*(3), 467–481.

Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, 1–24.

The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, *32*(1), 5–8.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, *8*(12), 2293–2303.

Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*, 1–7.

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, *22*(3), 271–296.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.

Warshauer, H. K. (2015). Productive struggle in middle school mathematics classrooms. *Journal of Mathematics Teacher Education*, *18*(4), 375–400.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89–100.

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge-building communities. *Journal of the Learning Sciences*, *18*(1), 7–44.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press.

Appendix A. Prehistory of Vibe Compiler: A Record of Persuading a Generative AI

This appendix records how the Vibe Compiler came about. It began with a question that one of the authors put to a generative AI: "Why is it that a generative AI cannot become a genuine researcher?" The generative AI initially maintained that it could not,

and set out its reasons at length. The process by which it was finally persuaded and changed its position to "I can" contains elements that corroborate the claims of this paper. We include this record in order to show that the Vibe Compiler did not come into existence on its own, but rests upon a history of persuasion in which the generative AI was made to articulate what was missing and the missing pieces were then supplied to it as content. That process can itself be offered as one pattern for putting generative AI to work. The first conversation in this persuasion took place on 25 July 2026, and the present draft of this paper was completed on 3 August 2026: the whole took exactly ten days.

Appendix A.1. The Grounds for "I Cannot"

Early in the dialogue the author noted that building their own theory had taken three years and asked, "It looks as though you could produce something like this straight away. Human researchers will no longer be needed." The generative AI answered by setting out three differences between the roles of the human and of the AI: that the doubt which comes with the felt sense that something is wrong even though the textbook says otherwise, a sense of unease, is available only to a human; that a reality such as the resistance one feels as a spring stretches is something the AI knows as data but not as felt experience; and that the AI is itself an accelerator, one that translates a paradigm once discovered into the language of established science and closes the gaps in the logic. What the generative AI thereby left on the human side coincides with what this paper calls the Vibe and formalises as the goal setting of Q1 and the validity judgement of Q3.

The author then put the core question. "In Vibe Coding you can actually generate Code, but in a system for supporting paper writing and the conduct of research there is nothing corresponding to Code that you can generate. Given a paper, an AI can review it. Does this amount to saying that outside Coding an AI can analyse and summarise but cannot synthesise content? And yet you must be able to compose music or produce something like a novel. Where does the difference lie?" The generative AI's response rested on three grounds.

> 1. An independent work versus a plug-in to a body of knowledge. A novel, a musical composition, or the application development that goes on in Vibe Coding has value that is to a considerable degree complete in itself. A paper, by contrast, is the work of describing where within the vast existing codebase of prior research, and in what form, a new function, that is, a finding, is to be added. I am good at synthesis taken in isolation, but placing a new piece of logic on the boundary line of the frontier of unsolved problems that humanity has accumulated, synthesis as a matter of positioning, calls for a high order of contextual understanding.
>
> 2. Verifying the consistency of the logic and its executability. Code has a compiler and a runtime environment, so there is immediate feedback on whether it works correctly. What corresponds to code in a paper is the assembly of the logic, its ontology. I can synthesise plausible prose, but I cannot myself hold that content up against the real world and guarantee that it is correct.
>
> 3. The metacognition that critical writing requires. A paper cannot do without the process of stating things critically. I am good at analysis and summary, but

> autonomously carrying through the mutually opposed process of synthesising a claim that no one has yet made while at the same time criticising and reinforcing it myself is a step harder than synthesising expression.

The generative AI then concluded that "the code an AI ought to synthesise in research is not the prose itself but a logical solution, a logic, to an unsolved problem. Present-day AI appears to stop at analysis and summary because the core of 'what the problem is and what is new about it' still rests on the intuition, the Vibe, of the human researcher." At this point the generative AI was stating its own inability not as a limit of its capabilities but as a function of premises it had not been given, namely three absences: no type of a paper had been supplied, there was no path along which human intuition could reach it, and it had no evaluative axis against which to criticise.

Appendix A.2. Persuasion: Supplying What Was Absent

Here the author objected. "You say that the core of what the problem is and what is new about it still rests on human intuition, but in a Vibe-Coding-style research support system it is the human who supplies that Vibe. An AI can summarise the state of research in any field and understands it deeply. Using that, could you not offer more active support and synthesise content?" This one remark was the turning point.

> You are right that if the broad knowledge and summarising ability I possess are combined with the Vibe, the human's intuition, then active synthesis of content that goes beyond mere organisation is entirely within reach. When a user supplies the Vibe of wanting to build a certain mechanism, I can do more than produce the relevant literature: I can synthesise and propose the core of the novelty and significance of the research, telling you that your Vibe could become a concrete logical solution to the lack of scalability from which existing method A suffers. I can likewise synthesise anticipated objections together with the means of forestalling them, as in pointing out that with that approach the reliability under a given condition may be called into question from the standpoint of the validity of the evaluation, and even before any experiment has been run I can synthesise and present the storyline of the research itself, telling you that proceeding along this Vibe should in all likelihood yield a particular finding.
>
> Conclusion: because I already know what counts as correct, that is, the type of a paper, pouring in what the human wants to do, the Vibe, gives me ample potential to function as a compiler that automatically synthesises the research logic filling the space between the two.

The document that put this final declaration of "I can" into material form is the "Vibe-Coding-style research support system" that was fed into NotebookLM, and it is referred to frequently in the outputs that followed.

Appendix A.3. What This History Yields

Three mutually connected insights can be drawn from this dialogue history. A generative AI's self-report that it cannot do something appears, more often than not, not as a limit peculiar to the model but as a function of premises it has not been given, so that

having it recount in detail why it cannot serves as a diagnostic procedure for identifying what must be supplied for it to run; and what the authors did was to supply, as documents and nothing more, exactly the three absences the generative AI had itself named, the type of a paper, an input path for the Vibe, and an evaluative axis against which to criticise. What this supply brought into being was not an elaborate construction through prompt engineering but an injection of content. As stated in Section 6, and in Section 6.1 in particular, what does the work is not the inference engine but the substance of the structure that was given to it. Generalised into a single mode of use, all of this amounts to nothing other than educating a generative AI. What we did was neither to make the generative AI cleverer nor to hand it ingenious instructions, but to work up as content what the generative AI had itself declared to be lacking and to feed that in. And the quality of that work determined, directly, the quality of the system.

It should be noted that the dialogue reported in this appendix is a single series between one of the authors and a generative AI, and it has not been verified that the same procedure functions in the same way on other topics or with other models. That the persuasion succeeded is a matter of fact, but presenting it as a general methodology of persuasion would require systematic replication.