Educational Innovation Research, 2026, Volume 4, Issue 7
https://bbwpublisher.com/journal/EIR
Online ISSN: 3029-1852
Print ISSN: 3029-1844

Grey–Copula Dynamic Prediction of Hidden Learning Risk in Vocational Mathematics under Generative AI

Yan Wang*

School of Humanities, Sichuan Vocational College of Information Technology, Guangyuan 628000, Sichuan, China

Author to whom correspondence should be addressed.

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Generative AI can raise students' assignment marks while concealing a decline in their unaided problem solving. In vocational college mathematics, a semester supplies only twelve to sixteen weekly observations per student, so large-sample classifiers cannot be estimated reliably; risk appears as several indicators deteriorating together, so single-threshold rules overlook students approaching two limits at once; and dependence between indicators shifts once generative AI enters the classroom, so static models respond too slowly. This paper proposes a grey–Copula dynamic prediction method built around these three constraints. Four weekly indicators operationalise hidden learning risk as two conjunctive events; an improved grey model produces one-step forecasts and threshold-crossing probabilities from a few weeks of data; a rolling-window Copula couples paired indicators and is re-estimated weekly; and joint probabilities are mapped onto a three-tier alert. In a simulation of 15 scenarios and 1,600 replications, the joint probabilities were better calibrated than threshold and single-indicator rules (Brier score 0.115 against 0.177 and 0.262), and the rolling window recovered the post-break dependence within about 2.5 weeks. The Copula layer improved recall more than calibration, and an autoregressive alternative forecast better. No classroom validation is claimed.

Keywords: Generative artificial intelligence; Vocational college mathematics; Learning risk; Grey prediction model; Copula function

Online publication: July 26, 2026

1. Introduction

Generative artificial intelligence (GenAI) has become an ordinary tool in mathematics classrooms. A student photographs a problem, receives a complete solution within seconds, and copies a route to the answer that looks like understanding. In vocational colleges, where many students enter with weak secondary-school mathematics, this convenience meets a deliberately procedural curriculum — standard problem types, clear steps, predictable marks — and these are the features GenAI reproduces most easily.

Short-run performance is a poor guide to what has been learned. Unrestricted access to GPT-4 raised practice scores by 48%, yet students who had used the tool scored 17% below those who had never used it once it was withdrawn [1]. A randomised study of university students' writing reached a parallel conclusion: assisted essays earned higher marks while knowledge gain and transfer did not differ, a pattern the authors name metacognitive laziness [2]. The mechanism is cognitive offloading; when the reasoning is delegated, the practice that would build it never takes place, and the deficit stays invisible for as long as the tool is present.

What makes the deficit hard to manage is its timing. Assignment accuracy stays normal, the loss surfaces in an unaided quiz, and by then the midterm has passed. A weekly signal would help, but three constraints stand in the way. A semester yields twelve to sixteen observations per student, too few to estimate a large-sample classifier. Risk is carried by co-deterioration, so indicator-by-indicator thresholds miss the student whose help-seeking rises while transfer falls. And the dependence between indicators is not stable: once GenAI becomes routine, help-seeking and transfer move together more or less strongly, and a model estimated once cannot follow it.

The paper defines the construct (Section 2), builds the model (Section 3), reports a simulation (Section 4), sets out the application pathway (Section 5), and discusses and concludes (Sections 6–7).

2. Hidden learning risk and its measurement

2.1. A gap between assisted and unassisted performance

We define hidden learning risk as the state in which a student can complete a task with AI assistance but cannot complete an isomorphic task without it. Using GenAI is not in itself a risk indicator: a student who consults a tool, understands the explanation and then transfers the method is using the tool as a scaffold. Risk lies in the gap between assisted and unassisted performance, and the gap is hidden because the assisted performance is what teachers observe, mark and record. The construct therefore has to be measured twice, once under assistance and once without it, on a task of the same structure but a different surface.

2.2. Four weekly indicators

Four indicators carry the construct, all proportions in [0, 1] and all collectible within the ordinary timetable (**Table 1**). Independent accuracy (IA) comes from a short, time-limited set solved with AI switched off. Repeated-error rate (RE) comes from tagging homework errors against a cumulative list of error types and counting the share that repeat an earlier type. Help-seeking dependence (AD) records how often a student turns to the tool per task, checked by an unaided retest. Transfer accuracy (TA) comes from an isomorphic item attempted immediately after an AI explanation.

Table 1. Weekly indicators of hidden learning risk

Indicator	Symbol	Weekly collection	Direction and threshold
Independent accuracy	IA	Correct share on 3–5 items solved without AI, under time limit	< 0.60
Repeated-error rate	RE	Share of this week's errors repeating earlier error types	> 0.50
Help-seeking dependence	AD	AI queries per task, checked by an unaided retest	> 0.70
Transfer accuracy	TA	Correct share on isomorphic new items after an AI explanation	< 0.60

Thresholds are project-set values and require calibration in a real classroom. Source: compiled by the author.

2.3. Risk as a conjunction, not a threshold

No single indicator crossing its threshold is treated as risk: a student who seeks help often but transfers well is being scaffolded. Risk is therefore a conjunction. Event E_1 = {AD > 0.70 and TA < 0.60} describes dependence that has not been converted into competence, and event E_2 = {RE > 0.50 and IA < 0.60} describes the absence of effective error correction. The composite event is $U = E_1 \cup E_2$, and the target is the probability that an event occurs in week n + 1 given the observations up to week n. A probability can be ranked when time is short, and it avoids the false precision of a single-indicator label.

3. A grey–Copula dynamic prediction model

3.1. Architecture

The model has four layers: data (four indicators, weekly), marginal (one-step forecasts converted into threshold-crossing probabilities), dependence (paired indicators on a rolling window) and output (joint risk and alert level).

3.2. An improved grey model for short series

Grey prediction suits the situation a semester presents: a short series, no reliable distributional assumption, and a need to extrapolate one step rather than explain a population [5]. The classical model accumulates the series to weaken randomness, fits a first-order differential equation and recovers the forecast by inverse accumulation.

First, the indicators lie in [0, 1] and a grey forecast can leave that interval, so a shifted logit is applied before modelling and inverted afterwards:

$$y_k = \ln[p_k/(1 - p_k)] + 5, \quad p_k = (x_k + 0.02)/1.04 \tag{1}$$

where x_k is the observed proportion in week k and the shift keeps y_k positive. Second, classical accumulation weights early and recent weeks alike, so fractional-order accumulation with order $r \in (0, 1]$ is used to give recent weeks more weight:

$$y^{(r)}(k) = \Sigma_{i=1}^{k}\, C(k - i + r - 1, k - i)\, y(i) \tag{2}$$

where C is the generalised binomial coefficient defined through the gamma function; $r = 1$ returns the classical model. Third, because a plain average background value biases the estimate when the series bends, a weight λ is introduced,

$$z^{(r)}(k) = \lambda\, y^{(r)}(k) + (1 - \lambda)\, y^{(r)}(k - 1), \quad \lambda \in [0.3, 0.7] \tag{3}$$

so that $y^{(r)}(k) - y^{(r)}(k - 1) + a\, z^{(r)}(k) = b$ can be estimated by least squares, with a the development coefficient and b the grey action; writing B for the matrix of rows $(-z^{(r)}(k), 1)$ and Y for the vector of first differences, the estimate is $(\hat{a}, \hat{b})^T = (B^T B)^{-1} B^T Y$. Fourth, the response function

$$\hat{y}^{(r)}(k + 1) = [y^{(r)}(1) - \hat{b}/\hat{a}]\, e^{-\hat{a}k} + \hat{b}/\hat{a} \tag{4}$$

gives the accumulated forecast, and a residual correction absorbs persistent bias, $\tilde{y}(n + 1) = \hat{y}(n + 1) + \varphi\, e(n)$ with $e(n) = y(n) - \hat{y}(n)$ and $\varphi \in \{0, 0.2, \ldots, 1\}$. The settings (w, r, λ, φ) minimise the mean absolute error over the four most recent origins and are shared by the class, so a student contributes a few weeks of observations rather than a training set; if the best configuration beats the classical model by less than 2%, the

classical model is kept.

3.3. From forecast to marginal risk

A point forecast is not an alert. The probability that an indicator will cross its threshold next week is read from the empirical distribution of recent residuals rather than from a normal assumption that a few dozen residuals cannot support. For help-seeking dependence, with E the pooled residuals of the last four forecast origins, c the threshold and $g(\cdot)$ the transformation in equation (1),

$$p_{AD} = (\#\{e \in E : e > g(c) - g(\tilde{y})\} + 0.5)/(|E| + 1) \tag{5}$$

with the inequality reversed for indicators whose risk lies below the threshold. The half-count and the unit in the denominator keep the probability away from zero and one; the cost is a coarse scale, with about 2.5 percentage points at forty residuals.

3.4. Dependence through a rolling-window Copula

Marginal probabilities treat the indicators as independent. Copulas separate dependence from the margins: by Sklar's theorem the joint distribution of a pair of residuals is a Copula on the two margins [3,4]. With pseudo-observations oriented so that smaller is worse, the probability of the conjunctive event is

$$P(E_1) = C(p_{AD}, p_{TA}; \theta) \tag{6}$$

where θ parameterises dependence and independence corresponds to $C(u, v) = uv$. Seven families are considered; because simultaneous deterioration is the event of interest, lower-tail dependence is the relevant property, and for the Clayton family $C(u, v; \theta) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$ with lower-tail coefficient $\lambda_L = 2^{-1/\theta}$. Parameters are estimated by pseudo-maximum likelihood on a four-week rolling window, with residuals demeaned by week and ranked, and the Akaike information criterion chooses the family. The short window is a behavioural choice: a long window would average the new regime with the old.

3.5. Joint risk and three-tier alerts

The two events involve disjoint pairs of indicators and their mutual dependence cannot be estimated from four weeks of data, so they are treated as conditionally independent,

$$P(U) = 1 - [1 - P(E_1)][1 - P(E_2)] \tag{7}$$

and the alert level follows from the joint probability,

$$L(P) = 1 + I(P \geq 0.30) + I(P > 0.60) \tag{8}$$

where $I(\cdot)$ is the indicator function and $L = 1, 2, 3$ denotes low, moderate and high risk. The cut-points are not estimates but administrative choices about how many students a teacher is asked to follow up in a given week, and they should be set with intervention capacity in mind.

4. Simulation

4.1. Design

Because thresholds and dependence are stipulated, the method was evaluated on simulated data in which true risk is known. A class of 120 students was followed for 16 weeks: weeks 1–8 warm-up, weeks 9–16 predicted and scored. Students were drawn from three profiles, stable (60%), AI-dependent (20%) and volatile (20%), and each indicator was generated as an individual mean plus a learning curve $0.35(1 - e^{-0.15t})$ plus an AR(1)

disturbance, observed as a binomial proportion over 20 items, retaining the noise of real marking. GenAI arrived in week 10; help-seeking then rose across the class, and transfer and error correction fell for dependent students. The dependence between AD and TA disturbances came from a Clayton Copula with τ rising from 0.30 to 0.60 at the break; RE and IA kept τ = 0.30. In the baseline scenario, E_1 occurred in 17.4% of student-weeks and U in 21.9%.

Fourteen further scenarios varied class size (60 and 240), term length (12 and 20 weeks), post-break dependence (τ = 0.45 and 0.75), the Copula family (Frank, t), the timing of the break (none, week 13), item count (10 and 40), independence of the shocks, and a class-wide trend. The baseline was replicated 200 times and the other scenarios 100 times, 1,600 replications in all with fixed seeds.

4.2. Point forecasts

The improved grey model (IGM) was compared with a naive forecast, GM(1,1), a metabolic GM(1,1) with a four-week window, and a panel AR(1). The IGM had the lowest mean absolute error of the three grey variants (0.1509 against 0.1591 and 0.1898; **Table 2**) and beat GM(1,1) in every replication and scenario after Holm correction. Two results qualify the claim: the IGM was only 0.0023 better than the naive forecast, and the panel AR(1) was better still (0.1370) in every scenario, because the AR model is correctly specified for the generating process. With strongly autocorrelated real data, the grey model may not be the best predictor; it needs no estimated autocorrelation structure, starts from a few weeks of data, and cannot forecast outside the admissible range.

Table 2. One-step forecast error, baseline scenario (mean over 200 replications)

Method	MAE	RMSE	MAPE (%)
Naive (last week's value)	0.1532	0.1991	40.11
GM(1,1)	0.1591	0.2040	42.96
Metabolic GM(1,1), w = 4	0.1898	0.2451	49.77
Improved grey model (IGM)	0.1509	0.1960	38.73
Panel AR(1)	0.1370	0.1746	37.85

Four indicators pooled. Standard deviations are below 0.004 for MAE and RMSE and about 1.2 for MAPE. Source: simulated by the author.

4.3. Joint risk

For E_1, the joint probabilities were better calibrated and more discriminating than the rules they replace (**Table 3**): the Brier score was 0.1150 against 0.1767 for a threshold rule applied to the same point forecasts and 0.2615 for a single-indicator rule on AD, with AUC 0.802 against 0.682 and 0.692; the paired differences were significant after Holm correction in all 15 scenarios. The gain comes from issuing a probability rather than a cut: students near a threshold receive intermediate probabilities instead of being treated alike. Two comparisons are less favourable. Logistic regression matched the Brier score (0.1157) with a higher AUC, but much lower recall (0.381), and it was better in the no-break, late-break, and 12-week scenarios. The Copula layer added little to calibration (0.1150 against 0.1145 under independence) but raised recall from 0.526 to 0.576 and F1 from 0.494 to 0.499, and with GM(1,1) margins the Brier score rose to 0.1233: the dependence correction is only as good as the margins.

Table 3. Alert performance for E1, baseline scenario (mean over 200 replications)

Method	Brier	AUC	Recall	F1
Grey–Copula, rolling window	0.1150	0.802	0.576	0.499
Independence (no Copula)	0.1145	0.803	0.526	0.494
Static Copula	0.1147	0.802	0.561	0.498
GM(1,1) margins + Copula	0.1233	0.772	0.516	0.459
AR(1) margins + Copula	0.1143	0.807	0.423	0.470
Logistic regression	0.1157	0.805	0.381	0.449
Point-forecast threshold rule	0.1767	0.682	0.468	0.477
Single indicator (AD > 0.70)	0.2615	0.692	0.622	0.451

Recall, precision and F1 use $P \geq 0.30$ as the alert cut. Source: simulated by the author.

4.4. Tracking the shift in dependence

The rolling window was built for the week-10 break and tracked it: the estimated τ rose from about 0.14 to 0.24 and reached the post-break level in 2.5 ± 1.1 weeks, against 6.0 ± 1.9 weeks for an expanding window, which never caught up in 35% of replications against 0.5% for the rolling window (paired difference −3.4 weeks). Faster tracking did not translate into a lower Brier score on grey margins. Tail dependence was underestimated: a lower-tail family was selected in only about 30% of runs and the lower-tail coefficient was off by 0.542, against 0.561 and 0.595 for expanding-window and Gaussian-only specifications.

The alert bands behaved monotonically, with observed E_1 incidence of 0.094, 0.362 and 0.653 in the low, moderate and high bands. AR margins discriminated more steeply (0.115, 0.500, 0.797) but flagged only 28% as many student-weeks as high risk, which is less useful where follow-up capacity binds. None of this shows that the method improves learning: it shows calibrated probabilities when the generating mechanism matches the model's assumptions, not that it would in a classroom.

5. Running the method in an ordinary teaching week

The data come from work a teacher already does. In the last ten minutes of the week's final lesson, students complete a short set with AI switched off: three to five routine items, which give IA, and one item isomorphic to a problem explained with AI earlier in the week, which gives TA. Homework errors are tagged during marking, which gives RE, and queries to the school-approved platform give AD. Weeks 1–8 fix the parameters; alerts begin in week 9, leaving half the semester for intervention. Because parameters are shared, a class of forty is modelled as forty short series.

Interventions are matched to the band and kept short. Moderate risk receives an error-feedback sheet listing the error types that recurred in the previous fortnight. High risk adds a timed unaided redo, in which the student reworks a problem originally completed with assistance and explains the key step aloud. Students flagged on E_1 receive an isomorphic transfer item immediately after the next AI explanation, so that comprehension is converted into performance before the lesson ends.

Three boundaries should be fixed before collection begins. Alerts support teaching only; they are not evidence for assessment, ranking or disciplinary action, and a mark must never depend on a predicted probability. Results are visible to the teacher and to the student concerned, and feedback is phrased as a possibility rather than a judgement. Only the four indicators are stored, records are deleted at the end of

the semester, and students are told plainly that the model watches whether help turns into independent performance, not whether help was sought.

6. Discussion

Three design decisions carry the argument. Reading marginal probabilities from residuals rather than a normal curve keeps them transparent when a teacher asks why a student was flagged. Defining risk as a conjunction encodes the joint nature of the phenomenon and yields a ranking rather than a label. Re-estimating dependence weekly follows from the claim that GenAI changes how indicators co-move; the simulation supports the responsiveness and shows its price, since four weeks contain few observations.

The vulnerabilities are equally clear. Errors in the marginal layer propagate into the joint probability, and the Copula can reweight margins but not repair them. Tail dependence estimated from short windows is underestimated, which biases conjunctive probabilities downward in the cases that matter most. Conditional independence understates risk when a student deteriorates across all four indicators at once. And the thresholds are administrative values that have not been calibrated against outcomes.

7. Conclusion

The characteristic risk of generative AI in vocational mathematics is not heavy use but a widening gap between assisted and unassisted performance, and that gap can be turned into a weekly signal with modest collection. The proposed method defines risk as a conjunction of indicators, forecasts each indicator with a grey model built for short series and bounded proportions, converts forecasts into threshold-crossing probabilities, and couples paired indicators through a rolling-window Copula. Simulation shows calibrated joint probabilities, a clear advantage over threshold and single-indicator rules, and fast recovery of a break in dependence, alongside a Copula gain that rests on the quality of the margins and an autoregressive alternative that forecasts better under matched assumptions.

No classroom validation is claimed: the thresholds are stipulated, conditional independence is an approximation, and the behaviour of the rolling window under real classroom noise is unknown. The next step is a one-semester study in cooperating vocational colleges that collects the four indicators under informed consent, calibrates thresholds against observed outcomes, and compares grey marginals with autoregressive alternatives, with parallel classes as controls. Until then, the method is a design for a signal a teacher could act on, not evidence that acting on it improves learning.

Disclosure statement

The author declares no conflict of interest.

References

[1] Bastani H, Bastani O, Sungu A, et al., 2025, Generative AI Without Guardrails Can Harm Learning: Evidence From High School Mathematics. PNAS, 122(26): e2422633122.

[2] Fan Y, Tang L, Le H, et al., 2025, Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence

on Learning Motivation, Processes, and Performance. British Journal of Educational Technology, 56(2): 489–530.

[3] Sklar A, 1959, Fonctions de Répartition à N Dimensions et Leurs Marges. Publications de l'Institut de Statistique de l'Université de Paris, 8: 229–231.

[4] Nelsen RB, 2006, An Introduction to Copulas. 2nd ed. New York: Springer.

[5] Deng JL, 1982, Control Problems of Grey Systems. Systems & Control Letters, 1(5): 288–294.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.