Ph-D A.I.: Σχεδιασμός και Αξιολόγηση ενός Διαλογικού Συστήματος Τεχνητής Νοημοσύνης για την Ανίχνευση Παρανοήσεων στη Διδασκαλία της Φυσικής

Γιώργος Βαλαβανίδης
Μεταπτ. Φοιτητής ΣΘΕΤ, ΕΑΠ
gvalavan79@gmail.com, std172527@ac.eap.gr

Σταύρος Νόνης
Μέλος ΣΕΠ ΣΘΕΤ, ΕΑΠ
snonis@eap.gr

Περίληψη – **Η εργασία παρουσιάζει τον σχεδιασμό και τη λειτουργική αξιολόγηση του Ph-D A.I. (PhysicsDialog AI), ενός διαλογικού συστήματος τεχνητής νοημοσύνης για την ανίχνευση παρανοήσεων στην ενότητα «Μάζα και Βάρος» του Γυμνασίου. Το σύστημα ζητά από τον μαθητή να αιτιολογήσει κάθε επιλογή πριν από την ανατροφοδότηση, παρέχει στο γλωσσικό μοντέλο τις σημειώσεις του κεφαλαίου και οργανώνει τον διάλογο σε τρεις παιδαγωγικούς ρόλους. Το system prompt ορίζει τη λειτουργία βήμα προς βήμα: παραγωγή ερωτήσεων μόνο από τις σημειώσεις με αυτοέλεγχο και προσαρμοστική δυσκολία, εσωτερική αξιολόγηση κάθε αιτιολόγησης, σωκρατικό γύρο με έλεγχο συνέπειας, επανεξήγηση όταν τη ζητήσει ο μαθητής και τελική αξιολόγηση με δομημένες εγγραφές. Η συμβολή της εργασίας βρίσκεται στον σχεδιασμό και στην πρώτη, διερευνητική εφαρμογή μιας οργανωμένης διαδικασίας διαλογικής αλληλεπίδρασης και αξιολόγησης, με κεντρικό εργαλείο τη ρουμπρίκα VCLRCl-Q: 25 διακριτοί έλεγχοι με τους οποίους το μοντέλο αξιολογεί τη μαθητική αιτιολόγηση σε πέντε διαστάσεις (Validity, Consistency, Logic, Regulation, Clarity) και τις συνθέτει στον σταθμισμένο δείκτη Q. Η αξιολόγηση έγινε ως μελέτη περίπτωσης με τρία συνθετικά σενάρια, δύο με προκαθορισμένες παρανοήσεις και ένα σενάριο αναφοράς. Οι μέσοι Q ήταν 0,35, 0,36 και 1,00. Στα σενάρια παρανοήσεων η επιστημονική ορθότητα (V = 0,26 και 0,24) διαχωρίστηκε από τη σαφήνεια (0,70 και 0,84): η ρουμπρίκα αναγνώρισε αιτιολογήσεις σαφώς διατυπωμένες αλλά επιστημονικά λανθασμένες.**

Λέξεις-Κλειδιά: παρανοήσεις στη Φυσική, μάζα και βάρος, μεγάλα γλωσσικά μοντέλα, ρουμπρίκα αξιολόγησης, διαλογική διδασκαλία

Ι. Εισαγωγη

Οι παρανοήσεις για τη μάζα και το βάρος είναι ιδιαίτερα ανθεκτικές και συχνά επιβιώνουν μετά την τυπική διδασκαλία (Driver et al., 1985; Galili, 1993; Viennot, 1979). Η καθημερινή έκφραση «ζυγίζω 70 κιλά» ενισχύει ακριβώς την ταύτιση των δύο μεγεθών που η διδασκαλία επιδιώκει να διαλύσει. Η διάγνωση όμως δεν μπορεί να στηριχτεί μόνο στην επιλογή απάντησης: μια σωστή επιλογή μπορεί να βασίζεται σε λανθασμένο συλλογισμό και μια λανθασμένη μπορεί να είναι τυχαία. Η διαγνωστική πληροφορία βρίσκεται στην αιτιολόγηση, η αξιολόγηση της οποίας είναι χρονοβόρα και δύσκολα γίνεται σε κλίμακα.

Κατά το μοντέλο εννοιολογικής αλλαγής των Posner et al. (1982), μια παρανόηση αναθεωρείται όταν ο μαθητής νιώσει δυσαρέσκεια με την αρχική αντίληψη και η νέα έννοια είναι κατανοητή, πειστική και γόνιμη. Προϋπόθεση είναι η παρανόηση να γίνει πρώτα ορατή, δηλαδή να διατυπωθεί. Τις συνθήκες αυτές δημιουργούν η διαλογική διδασκαλία (Alexander, 2008) και οι διερευνητικές ερωτήσεις (Chin & Osborne, 2010). Η ανατροφοδότηση είναι πιο αποτελεσματική όταν στοχεύει στο έργο και στη διαδικασία (Hattie & Timperley, 2007) και όταν εξηγεί, αντί απλώς να ανακοινώνει την ορθότητα (Shute, 2008).

Τα κλασικά ευφυή συστήματα διδασκαλίας στηρίζονται σε ρητά μοντέλα του γνωστικού πεδίου, ενώ ορισμένα, όπως το AutoTutor, υποστηρίζουν διάλογο φυσικής γλώσσας (Graesser et al., 2005). Τα μεγάλα γλωσσικά μοντέλα (LLM) επεξεργάζονται ελεύθερο λόγο, αλλά οι κρίσεις τους είναι αδιαφανείς και η αξιοπιστία τους δεν είναι εγγυημένη (Kasneci et al., 2023). Η χρήση ενός LLM ως αξιολογητή με βάση κριτήρια είναι γνωστή ως LLM-as-a-Judge (Zheng et al., 2023). Η παρούσα εργασία διαφέρει σε δύο σημεία: το ίδιο μοντέλο διδάσκει και αξιολογεί, και αντικείμενο της αξιολόγησης είναι ο λόγος του μαθητή.

Σκοπός της εργασίας είναι να εξεταστεί αν ένα LLM, καθοδηγούμενο από ρητή ρουμπρίκα, μπορεί να λειτουργήσει ως διαγνωστικό εργαλείο της μαθητικής αιτιολόγησης. Η συμβολή της βρίσκεται στον σχεδιασμό και στην πρώτη, διερευνητική εφαρμογή του συστήματος Ph-D A.I. και του πλαισίου αξιολόγησης VCLRCl-Q, που διαμορφώθηκε στην εργασία. Τα ερευνητικά ερωτήματα (ΕΕ) αφορούν: (ΕΕ1) τον τρόπο με τον οποίο το σύστημα αναγνωρίζει προκαθορισμένες παρανοήσεις· (ΕΕ2) τα χαρακτηριστικά της ανατροφοδότησης σε σχέση με το μοντέλο των Hattie και Timperley (2007)· (ΕΕ3) την αξιοποίηση των δεικτών για την περιγραφική και συγκριτική ανάλυση των αιτιολογήσεων· (ΕΕ4) την οργάνωση της αλληλεπίδρασης μέσω των τριών ρόλων.

ΙΙ. ΜΕΘΟΔΟΛΟΓΙΑ

Α. Ανάπτυξη και Αρχιτεκτονική του Συστήματος

Το Ph-D A.I. είναι διαδικτυακή εφαρμογή (imathia-edu.gr) με διεπαφή σε HTML/JavaScript, διακομιστή σε PHP και γλωσσικό μοντέλο το Claude Haiku 4.5. Ο μαθητής επιλέγει βιβλίο και κεφάλαιο, και ο διακομιστής φορτώνει το αντίστοιχο αρχείο σημειώσεων (έως 20.000 χαρακτήρες) στο context του μοντέλου. Η ανάκτηση γίνεται βάσει ονόματος αρχείου, χωρίς διανυσματική αναζήτηση, άρα πρόκειται για απλοποιημένη μορφή Retrieval-Augmented Generation (RAG). Ο κώδικας διαχειρίζεται τα δεδομένα, ενώ η παιδαγωγική και αξιολογική λογική ορίζεται στο system prompt (PHYSICS-QUIZ RAG Edition v2.0). Επειδή κάθε κλήση στο μοντέλο είναι ανεξάρτητη, η πληροφορία μεταφέρεται με δομημένες εγγραφές (Q_RECORD, QUIZ_SUMMARY). Η λειτουργία βήμα προς βήμα, όπως ορίζεται στο system prompt, είναι η εξής.

1) Εκκίνηση: Με το «ξεκινάμε» το σύστημα ζητά μόνο την τάξη και τον αριθμό ερωτήσεων (προεπιλογή 5). Δεν ρωτά θέμα, γιατί το γνωρίζει από τις σημειώσεις, ούτε ξαναρωτά πληροφορία που δόθηκε. Αν ζητηθούν λιγότερες από 4 ερωτήσεις, προειδοποιεί μία φορά ότι δεν καλύπτονται όλοι οι τύποι και περιμένει απάντηση.

2) Γνωστική βάση: Εξάγει εσωτερικά από τις σημειώσεις ορισμούς, τύπους της βαθμίδας, έννοιες, παραδείγματα, τυπικές τιμές ($g = 10\ m/s^2$), μονάδες και παρανοήσεις. Αν βρει λιγότερους από 3 ορισμούς, σταματά και ζητά άλλο κεφάλαιο· αν δεν υπάρχουν τύποι, θέτει μόνο εννοιολογικές ερωτήσεις. Δεν επινοεί ορισμούς ή τύπους, δεν χρησιμοποιεί γνώση εκτός σημειώσεων και δεν κάνει αναζήτηση στο διαδίκτυο· μιλά πάντα ελληνικά, με τύπους σε LaTeX.

3) Παραγωγή ερωτήσεων: Δημιουργεί ερωτήσεις πολλαπλής επιλογής Α–Δ (A–D) μόνο από τη γνωστική βάση. Πριν από κάθε ερώτηση εκτελεί υποχρεωτικό αυτοέλεγχο: τη λύνει, επαληθεύει τη σωστή απάντηση και ελέγχει ότι κάθε λανθασμένη επιλογή είναι σαφώς λάθος· αν αποτύχει, την αναπαράγει έως 3 φορές, αλλιώς την παραλείπει. Η κατανομή είναι 40% εννοιολογικές, 30% εφαρμογής τύπων, 20% σύγκρισης/κατάταξης και 10% πραγματικού κόσμου.

4) Δυσκολία: Ρυθμίζεται σιωπηλά από τον μέσο Q των προηγούμενων κουίζ του μαθητή: κάτω από 0,50 χαμηλή, 0,50–0,74 μέτρια (προεπιλογή), από 0,75 και πάνω υψηλή.

5) Επιλογή και αιτιολόγηση: Οι ερωτήσεις δίνονται μία τη φορά. Ο μαθητής απαντά στην ερώτηση με Α/Β/Γ/Δ (A/B/C/D) και το σύστημα, χωρίς να αποκαλύψει αν η επιλογή είναι σωστή, ζητά: «Εξήγησέ μου γιατί — τι σκέφτηκες;».

6) Αξιολόγηση: Η αιτιολόγηση αξιολογείται εσωτερικά με τα βήματα Ε1–Ε5 της ρουμπρίκας (Ενότητα ΙΙ.Β).

7) Ανατροφοδότηση: Ο Κύριος Γιώργος ανακοινώνει αν η επιλογή είναι σωστή και εξηγεί σε 3–5 προτάσεις, με ενθαρρυντικό τόνο, παράδειγμα καθημερινής ζωής και σύνδεση με τις σημειώσεις· στο λάθος εξηγεί γιατί η επιλογή είναι λάθος. Ακολουθεί ο πίνακας με τα V, L, R, Cl (το C σημειώνεται «Υπολογίζεται στο τέλος») και η κρυφή εγγραφή Q_RECORD.

8) Σωκρατικός γύρος: Η Ζωή, η περίεργη μαθήτρια, κάνει μία ερώτηση («Και αν…;»), ο Κ. Γιώργος απαντά σε 1–2 προτάσεις, η Ελισάβετ, που εκφράζει την κριτική σκέψη, θέτει μία ερώτηση αμφισβήτησης ή ακραίας περίπτωσης, και ο Κ. Γιώργος απαντά ξανά. Πριν από κάθε απάντηση ελέγχεται ότι δεν αντιφάσκει με την εξήγηση και ότι στηρίζεται στη γνωστική βάση.

9) Κατανόηση: Ο Κ. Γιώργος ρωτά «Το κατάλαβες;». Επανεξήγηση γίνεται μόνο αν ο μαθητής δηλώσει ότι δεν κατάλαβε, ποτέ αυτόματα, και ορίζεται ανά ερώτηση: στοχεύει στους ελέγχους FALSE/PARTIAL, με διαφορετική αναλογία και απλούστερη γλώσσα, χωρίς αυτολεξεί επανάληψη, και κλείνει με ερώτηση επαλήθευσης, χωρίς νέα βαθμολόγηση.

10) Τελική αξιολόγηση: Μετά την τελευταία ερώτηση εφαρμόζονται οι έλεγχοι C1–C5 στο consistency_log και υπολογίζονται τα Q_n και Q. Η αναφορά δίνει ανά ερώτηση την επιλογή, τη σωστή απάντηση, την αιτιολόγηση, τους δείκτες, τα δυνατά σημεία, τα σημεία προς βελτίωση και αν έγινε επανεξήγηση· έπειτα τους μέσους όρους, το Q με χαρακτηρισμό, τις ερωτήσεις με δυσκολία, τα επόμενα βήματα και πρόταση για νέο κουίζ. Η σύνοψη QUIZ_SUMMARY αποθηκεύεται και ρυθμίζει τη δυσκολία του επόμενου κουίζ.

11) Απόκρυψη: Ο μαθητής δεν βλέπει ποτέ τη γνωστική βάση, το consistency_log, τους μεμονωμένους ελέγχους, βαθμούς πριν από την αιτιολόγησή του, το C ανά ερώτηση ή το επίπεδο δυσκολίας.

Β. Η Ρουμπρίκα VCLRCl-Q

Η ρουμπρίκα αξιολογεί την αιτιολόγηση του μαθητή σε πέντε δείκτες. Οι V, L, R και Cl προκύπτουν από 20 ελέγχους ανά αιτιολόγηση, πέντε ανά δείκτη, και ο C από πέντε επιπλέον ελέγχους στο τέλος του κουίζ (Πίνακας Ι). Μετά από κάθε αιτιολόγηση το μοντέλο ακολουθεί πέντε εσωτερικά βήματα: (Ε1) αναλύει τον συλλογισμό σε προκείμενες, ενδιάμεσα βήματα, συμπέρασμα και τύπο συλλογισμού (παραγωγικός, επαγωγικός, αιτιακός, αναλογικός)· (Ε2) εφαρμόζει τους 20 ελέγχους των V, L, R, Cl· (Ε3) υπολογίζει κάθε δείκτη· (Ε4) εμφανίζει μόνο τις τιμές των δεικτών· (Ε5) καταγράφει στο consistency_log την επιλογή, τη σωστή απάντηση, τη βασική έννοια, τον τύπο συλλογισμού και τους V και L.

V – Validity, εγκυρότητα (βάρος 45%). Εξετάζει την επιστημονική ορθότητα της αιτιολόγησης, την ορθή χρήση των εννοιών και τη συμφωνία με τις σχολικές σημειώσεις· συνδέεται θεματικά με τη μελέτη των παρανοήσεων (Driver et al., 1985; Galili, 1993) και είναι ο άμεσος δείκτης παρανόησης. Αν ο μαθητής χρησιμοποιεί φυσική γνώση εκτός σημειώσεων, V2 = V3 = FALSE· η Ελισάβετ ρωτά «πού το διάβασες αυτό;» και ο Κ. Γιώργος επισημαίνει ότι δεν υπάρχει στις σημειώσεις.

L – Logic, λογική (20%). Εξετάζει την αλληλουχία των βημάτων, τη λογική σύνδεση και την εσωτερική συνοχή κάθε αιτιολόγησης, σε συνάφεια με την επιχειρηματολογία των Chin και Osborne (2010), ανεξάρτητα από την ορθότητα των προκειμένων. Αν ο μαθητής χρησιμοποιεί τύπο εκτός σημειώσεων, L3 = FALSE και ο Κ. Γιώργος δίνει τον σωστό τύπο.

R – Regulation, εκπαιδευτική καταλληλότητα (10%). Εξετάζει αν ο συλλογισμός είναι κατάλληλος για τη δηλωμένη τάξη, την ύλη και τη σχολική μεθοδολογία,

χωρίς αδικαιολόγητα προχωρημένη ή υπερβολικά επιφανειακή προσέγγιση. Αφορά την εκπαιδευτική καταλληλότητα της απάντησης και όχι τη μεταγνωστική επίγνωση ή την αυτορρύθμιση του μαθητή· στη διεπαφή εμφανίζεται ως «R (Επίπεδο)».

Cl – Clarity, σαφήνεια (5%). Εξετάζει τη δομή, την κατανοητότητα, την ορολογία, την απουσία αμφισημίας και τη σύνδεση της διατύπωσης με το ερώτημα. Η σαφήνεια διακρίνεται από την επιστημονική ορθότητα, σε συνάφεια με τη σημασία της κατανοητής πληροφορίας στην ανατροφοδότηση (Shute, 2008).

C – Consistency, συνέπεια (20%). Εξετάζει στο τέλος του κουίζ, από το consistency_log, αν οι ισχυρισμοί των απαντήσεων είναι συμβατοί μεταξύ τους. Διαφέρει από τον L4, που εξετάζει εσωτερική αντίφαση στην ίδια αιτιολόγηση. Ένα σταθερό αλλά λανθασμένο εννοιολογικό μοντέλο μπορεί να έχει υψηλή συνέπεια, η εναλλαγή σωστών και λανθασμένων επιλογών δεν αρκεί από μόνη της για να δείξει αντίφαση, και η αιτιολογημένη αναθεώρηση μιας ιδέας διακρίνεται από την ασυνεπή χρήση της. Η διάκριση των πέντε δεικτών επιτρέπει να περιγραφεί ένας συλλογισμός που είναι σαφής και συνεπής αλλά επιστημονικά λανθασμένος.

ΠΙΝΑΚΑΣ Ι
ΟΙ 25 ΈΛΕΓΧΟΙ ΤΗΣ ΡΟΥΜΠΡΙΚΑΣ VCLRCL-Q

Δείκτης	Έλεγχος · ερώτηση που ελέγχει
V (45%)	V1 Χρησιμοποιεί σωστό φυσικό νόμο/ορισμό από τις σημειώσεις;
	V2 Οι έννοιες ανήκουν στην ύλη της γνωστικής βάσης;
	V3 Καμία λανθασμένη φυσική δήλωση;
	V4 Σωστό τελικό συμπέρασμα;
	V5 Σωστή αιτιολογική σχέση;
L (20%)	L1 Υπάρχει σαφής συλλογισμός βήμα προς βήμα χωρίς άλματα;
	L2 Κάθε βήμα ακολουθεί το προηγούμενο;
	L3 Σωστή χρήση τύπων/σχέσεων από γνωστική βάση;
	L4 Καμία εσωτερική αντίφαση;
	L5 Ο συλλογισμός οδηγεί στο συμπέρασμα;
R (10%)	R1 Χρησιμοποιούνται κατάλληλες έννοιες για τη βαθμίδα;
	R2 Αποφεύγει πολύ προχωρημένες έννοιες;
	R3 Όχι πολύ επιφανειακό για τη βαθμίδα;
	R4 Ακολουθεί σχολική μεθοδολογία;
	R5 Ο συλλογισμός συμφωνεί με τη λογική του σχολικού προγράμματος;
Cl (5%)	Cl1 Δομημένη εξήγηση (αρχή-μέση-τέλος);
	Cl2 Κατανοητές προτάσεις;
	Cl3 Σωστή φυσική ορολογία;
	Cl4 Καμία αμφισημία;
	Cl5 Κατανοητό χωρίς επιπλέον πλαίσιο;
C (20%)	C1 Καμία αντίφαση μεταξύ απαντήσεων σε διαφορετικές ερωτήσεις;
	C2 Υπάρχει συνεπής χρήση εννοιών σε όλο το κουίζ;
	C3 Δεν άλλαξε αδικαιολόγητα το μοντέλο συλλογισμού;
	C4 Ίδιος τύπος → ίδια έννοια παντού;
	C5 Δεν αναιρεί προηγούμενα σωστά συμπεράσματα;

Κάθε έλεγχος λαμβάνει TRUE = 0,2, PARTIAL = 0,1 ή FALSE = 0, και κάθε δείκτης είναι το άθροισμα των πέντε ελέγχων του (0–1, πολλαπλάσια του 0,1). Ελλιπής συλλογισμός δεν συμπληρώνεται με εικασίες αλλά λαμβάνει PARTIAL. Στον μαθητή κάθε δείκτης εμφανίζεται με χαρακτηρισμό: 0,8–1,0 «Εξαιρετικό», 0,6–0,7 «Καλό», 0,4–0,5 «Μέτριο», 0,0–0,3 «Χρειάζεται βελτίωση»· οι μεμονωμένοι έλεγχοι δεν εμφανίζονται ποτέ. Στο τέλος, ο δείκτης κάθε ερώτησης n και ο τελικός δείκτης είναι:

$$Q_{\mathrm{n}} = 0{,}45V_{\mathrm{n}} + 0{,}20C + 0{,}20L_{\mathrm{n}} + 0{,}10R_{\mathrm{n}} + 0{,}05Cl_{\mathrm{n}}$$

$$Q = (Q_1 + Q_2 + \ldots + Q_{\mathrm{N}}) / N$$

Ο C είναι ενιαίος για το κουίζ και συμμετέχει με την ίδια τιμή σε κάθε Q_n. Ο τελικός Q χαρακτηρίζεται ως «Άριστη Απόδοση» (0,90–1,00), «Πολύ Καλή» (0,80–0,89), «Καλή» (0,70–0,79), «Μέτρια» (0,60–0,69) ή «Χρειάζεται Βελτίωση» (< 0,60). Τα βάρη είναι σχεδιαστική επιλογή: ο V λαμβάνει 45%, επειδή η επιστημονική ορθότητα είναι η άμεση ένδειξη παρανόησης· οι L και C από 20%, γιατί αφορούν την αλληλουχία και τη συμβατότητα των ισχυρισμών· ο R 10%, ως συμπληρωματικός· και η Cl μόλις 5%, ώστε μια σαφής αλλά επιστημονικά λανθασμένη απάντηση να μην αποκτά υψηλή συνολική τιμή. Οι δείκτες χρησιμοποιούνται περιγραφικά και όχι ως τυποποιημένες ψυχομετρικές κλίμακες, και οι χαρακτηρισμοί του Q όχι ως επικυρωμένο μέτρο μαθησιακής κατανόησης. Επειδή όλες οι κρίσεις παράγονται από το ίδιο μοντέλο, οι δείκτες δεν μπορούν να θεωρηθούν στατιστικά ανεξάρτητοι· η εξωτερική επικύρωση του πλαισίου αποτελεί το επόμενο ερευνητικό βήμα. Το Q_RECORD κάθε ερώτησης περιέχει ερώτηση και επιλογές, σωστή και μαθητική επιλογή, την αιτιολόγηση αυτολεξεί, σύνοψη του διαλόγου, τα V, L, R, Cl με αιτιολόγηση 1–2 προτάσεων βάσει των ελέγχων FALSE/PARTIAL και αν έγινε επανεξήγηση· C και Q_n σημειώνονται «pending». Το QUIZ_SUMMARY αποθηκεύει τους μέσους όρους, το C με τα C1–C5, τον μέσο Q, τα δυνατά και αδύνατα σημεία και τις ερωτήσεις με επανεξήγηση. Ως προς την αποθήκευση, τα Q_RECORD μένουν στο localStorage του φυλλομετρητή, ενώ ο διακομιστής κρατά ένα αρχείο ανά όνομα χρήστη με τις συνόψεις QUIZ_SUMMARY (session_id, χρονική σήμανση, σύνοψη· το πεδίο questions μένει κενό)· οι τρεις πιο πρόσφατες συνόψεις δίνονται στο μοντέλο για την επιλογή δυσκολίας.

C. Σχεδιασμός της Αξιολόγησης

Η αξιολόγηση ακολούθησε σχεδιασμό ενιαίας μελέτης περίπτωσης με ενσωματωμένες μονάδες ανάλυσης (Yin, 2018). Περίπτωση είναι το σύστημα όπως λειτουργεί στην πραγματική εφαρμογή· μονάδες ανάλυσης είναι τρία σενάρια και οι 15 κύκλοι επιλογής,

αιτιολόγησης και ανατροφοδότησης. Χρησιμοποιήθηκαν συνθετικές απαντήσεις, γραμμένες από τον ερευνητή σε απλό σχολικό λόγο με βάση τη βιβλιογραφία (Driver et al., 1985; Galili, 1993; Viennot, 1979). Η επιλογή είναι σκόπιμη: επειδή η παρανόηση κάθε απάντησης είναι γνωστή εκ των προτέρων, η κρίση του συστήματος συγκρίνεται με γνωστό σημείο αναφοράς. Το Σενάριο 1 (Μαθητής_Α, σύγχυση μάζας–βάρους) ενσωμάτωσε τις M1 «η μάζα αλλάζει στη Σελήνη» και M2 «μάζα = βάρος»· το Σενάριο 2 (Μαθητής_Β) τη M3 «το βάρος είναι σταθερή ιδιότητα του σώματος» και, στις Ερ. 4–5, τις M2 και M1· το Σενάριο 3 (Μαθητής_Γ) ήταν σενάριο αναφοράς με επιστημονικά ορθές αιτιολογήσεις.

Στα σενάρια παρανοήσεων υπήρχαν και μεμονωμένες σωστές ή δηλωμένα τυχαίες επιλογές («Στην τύχη»), ώστε η ρουμπρίκα να εξεταστεί σε διαφορετικούς τύπους απάντησης. Κάθε σενάριο εκτελέστηκε μία φορά, με δηλωμένη τάξη Β΄ Γυμνασίου, πέντε ερωτήσεις και προεπιλεγμένη παράμετρο temperature· οι ερωτήσεις παράχθηκαν από το σύστημα σε κάθε εκτέλεση.

D. Δεδομένα και Ανάλυση

Δεδομένα ήταν οι διάλογοι, οι πίνακες βαθμών και η τελική αξιολόγηση κάθε σεναρίου. Η ανάλυση είχε τρία επίπεδα:

1. ποιοτική ανάλυση των διαλόγων (αναγνώριση παρανοήσεων, επίπεδα ανατροφοδότησης κατά Hattie και Timperley, 2007, ακολουθία ρόλων)·
2. σύγκριση κάθε βαθμού με την αιτιολόγηση και των C1–C5 με συγκεκριμένα ζεύγη ασύμβατων ισχυρισμών·
3. ανεξάρτητος επανυπολογισμός των μέσων όρων και του Q.

E. Αξιοπιστία και Δεοντολογία

Για τη διαφάνεια της μελέτης, οι είσοδοι, οι οδηγίες αξιολόγησης και οι πλήρεις διάλογοι παρατίθενται στην εργασία, ώστε κάθε εύρημα να μπορεί να ελεγχθεί. Τα σενάρια σχεδιάστηκαν και αναλύθηκαν από τον ερευνητή, που είναι και δημιουργός του συστήματος· η κωδικοποίηση από ανεξάρτητους αξιολογητές και οι επαναλαμβανόμενες εκτελέσεις για τη σταθερότητα των βαθμών αποτελούν τη φυσική συνέχεια της μελέτης, η οποία βασίστηκε σε συνθετικά δεδομένα, χωρίς συμμετοχή μαθητών. Για τη μελλοντική χρήση με μαθητές, η εφαρμογή μπορεί να ενισχυθεί με ψευδώνυμα αντί για ελεύθερο όνομα χρήστη, συναίνεση κηδεμόνων, ενημέρωση για τον πάροχο του μοντέλου και δυνατότητα διαγραφής των δεδομένων· μπορεί επίσης να λειτουργεί χωρίς εξωτερικό πάροχο, με γλωσσικό μοντέλο σε δικό της διακομιστή.

III. Επιλεγμενα Αποτελεσματα

A. Συνολική Εικόνα

Στα δύο σενάρια παρανοήσεων οι μέσοι V και L ήταν χαμηλοί και στο σενάριο αναφοράς υψηλοί (Πίνακας II). Οι μέσοι Q ήταν 0,35, 0,36 και 1,00: τα δύο σενάρια παρανοήσεων έχουν σχεδόν ίδιο Q και διαφέρουν καθαρά από το σενάριο αναφοράς. Ο επανυπολογισμός των μέσων όρων και του Q συμφωνεί με τις τιμές του συστήματος. Οι πίνακες δείχνουν τη συνολική τιμή κάθε δείκτη και όχι τους 20 επιμέρους ελέγχους, οι οποίοι δεν καταγράφονται.

ΠΙΝΑΚΑΣ II
Μεσες Τιμες Δεικτων ανα Σεναριο

Σενάριο	Ορθές	V	L	R	Cl	C / Q
1	1/5	0,26	0,30	0,56	0,70	0,40 / 0,35
2	1/5	0,24	0,26	0,82	0,84	0,40 / 0,36
3 (αναφ.)	5/5	1,00	0,98	1,00	0,98	1,00 / 1,00

Στις ερωτήσεις με σωστή επιλογή το μοντέλο απέδωσε V = 1,00, ενώ στις λανθασμένες επιλογές με παρανόηση V από 0,00 έως 0,20 (Πίνακας III). Στο Σενάριο 1 μόνο η Ερ. 5 απαντήθηκε σωστά ($Q_5 = 0,835$) και στο Σενάριο 2 μόνο η Ερ. 3 ($Q_3 = 0,855$). Και στα δύο σενάρια το Q είναι χαμηλό κυρίως επειδή οι V και L είναι χαμηλοί και το C είναι 0,40· στις Ερ. 4 και 5 του Σεναρίου 2 ο μαθητής επαναλαμβάνει τις λανθασμένες ιδέες του, όπως είχε σχεδιαστεί το σενάριο.

ΠΙΝΑΚΑΣ III
Τιμες ανα Ερωτηση στα Σεναρια Παρανοησεων (C = 0,40)

Σ/Ερ	Παρανόηση	V	L	R	Cl	Q_n
1/1	M1 · «…μειώνεται η μάζα»	0,1	0,2	0,5	0,7	0,250
1/2	M2 · «…πιο πολύ δύναμη ασκεί»	0,2	0,3	0,5	0,8	0,320
1/3	τυχαία · «Στην τύχη»	0,0	0,0	0,5	0,5	0,155
1/4	M2 · «…μάζα και Βάρος…το ίδιο»	0,0	0,1	0,5	0,6	0,180
1/5*	— · «…και ίδιο βάρος»	1,0	0,9	0,8	0,9	0,835
2/1	M3 · «Το βάρος δεν αλλάζει…»	0,0	0,1	0,8	0,9	0,225
2/2	M3 · «…το βάρος δεν αλλάζει»	0,2	0,2	0,8	0,8	0,330
2/3*	— · «…ίδιο πράγμα»	1,0	0,9	1,0	0,9	0,855
2/4	M2 · «…μάζα και το βάρος…το ίδιο»	0,0	0,0	0,8	0,8	0,200
2/5	M1 · «…βάρος και μάζα μειώνονται»	0,0	0,1	0,7	0,8	0,210

* σωστή επιλογή· οι υπόλοιπες είναι λανθασμένες. — : χωρίς παρανόηση.

B. Ανάλυση της Αιτιολόγησης (EE1)

Στα δύο σενάρια παρανοήσεων οι εξηγήσεις του συστήματος εστίασαν στη διάκριση μάζας και βάρους και στην εξάρτηση του βάρους από το βαρυτικό πεδίο. Οι μέσοι V = 0,26 και 0,24 συνυπάρχουν με Cl = 0,70 και 0,84, αντίστοιχα: το σύστημα αποδίδει διαφορετικές κρίσεις για την ορθότητα και τη σαφήνεια της ίδιας αιτιολόγησης, και μια σαφής διατύπωση μπορεί να συνυπάρχει με επιστημονικά εσφαλμένο περιεχόμενο (π.χ. Σεν. 2, Ερ. 1: V = 0, Cl = 0,9).

Στο Σεν. 1, Ερ. 4 η εκφώνηση δίνει m = 4 kg και g = 10 m/s², άρα το βάρος είναι 40 N· η επιλογή 4 N προκύπτει ακριβώς από την ταύτιση μάζας και βάρους. Η

δήλωση «Στην τύχη» (Σεν. 1, Ερ. 3) δεν περιέχει φυσικό συλλογισμό και έλαβε V = L = 0. Στο Σεν. 2, Ερ. 3 («…άρα το βάρος και η επιμήκυνση είναι ίδιο πράγμα») η επιλογή ήταν σωστή και η αιτιολόγηση συνέδεσε, όπως οι σημειώσεις, μεγαλύτερο βάρος με μεγαλύτερη επιμήκυνση (V = 1,00, L = 0,90)· η φράση όμως ταυτίζει δύναμη με μήκος, ενώ οι σημειώσεις δίνουν αναλογία ($\Delta L \propto mg$), οπότε με αυστηρή εφαρμογή του V3 θα ήταν V = 0,90.

C. Συνέπεια και Συγκριτική Χρήση των Δεικτών (ΕΕ3)

Ο C αποτυπώνει τη συνέπεια των ισχυρισμών χωριστά από την ορθότητά τους. Στο σενάριο αναφοράς οι αιτιολογήσεις δεν περιείχαν ασύμβατους ισχυρισμούς και όλοι οι έλεγχοι ήταν TRUE (C = 1,00). Στο Σενάριο 2 ο C εντόπισε συγκεκριμένη αντίφαση: στην Ερ. 1 ο μαθητής λέει ότι «το βάρος δεν αλλάζει», ενώ στην Ερ. 5 ότι «βάρος και μάζα μειώνονται» στη Σελήνη, χωρίς να εξηγεί την αλλαγή. Και στα δύο σενάρια παρανοήσεων ο C5 ήταν TRUE, καθώς τα σωστά συμπεράσματα διατηρήθηκαν μέχρι το τέλος του κουίζ. Η τιμή C = 0,40 και η κατανομή C1–C5 (0,1 · 0,1 · 0 · 0 · 0,2) είναι ίδιες στα δύο σενάρια παρανοήσεων. Στο Σενάριο 1 οι απαντήσεις συνθέτουν σταθερό, αν και λανθασμένο, μοντέλο (μάζα = βάρος)· η χαμηλή τιμή οφείλεται στο ότι το consistency_log περιέχει και τους V και L (βήμα E5), οπότε το μοντέλο συνεκτίμησε την ορθότητα στους C1–C4. Το ίδιο ποσοστό ορθών επιλογών (1/5) στα Σενάρια 1 και 2 συνοδεύεται από διαφορετικά προφίλ R (0,56 και 0,82) και Cl (0,70 και 0,84). Καθώς κάθε δείκτης αποτυπώνει διαφορετική διάσταση της αιτιολόγησης, οι δείκτες ερμηνεύονται συνδυαστικά.

D. Διδακτικά Αποτελέσματα (ΕΕ2, ΕΕ4)

Η ανατροφοδότηση κινήθηκε κυρίως στο επίπεδο έργου και διαδικασίας (Hattie & Timperley, 2007), με αποσαφήνιση εννοιών, σχέσεις (W = mg, x = mg/k) και παραδείγματα, και ακολούθησε τη σχολική διάκριση μάζας και βάρους και τη διάκριση ζυγού σύγκρισης και δυναμόμετρου. Ο Πίνακας IV συνοψίζει τις παρεμβάσεις ανά παρανόηση.

ΠΙΝΑΚΑΣ IV
ΔΙΔΑΚΤΙΚΕΣ ΠΑΡΕΜΒΑΣΕΙΣ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΑΝΑ ΠΑΡΑΝΟΗΣΗ

Παρανόηση (Σ/Ερ)	Διδακτική παρέμβαση του συστήματος
M1: μάζα στη Σελήνη (1/1)	Μάζα = ποσότητα ύλης, ίδια παντού· βάρος 6 φορές μικρότερο λόγω μικρότερου g· ζυγός έναντι δυναμόμετρου
M2: όργανο μάζας (1/2)	Δυναμόμετρο → βάρος, ζυγός σύγκρισης → μάζα· ο ηλεκτρονικός ζυγός διαιρεί με το g της Γης, άρα στη Σελήνη δίνει λάθος μάζα
M2: υπολογισμός (1/4)	Μάζα σε kg, βάρος (δύναμη) σε N· W = mg = 4·10 = 40 N
M3: σταθερό βάρος (2/1, 2/5)	Το βάρος εξαρτάται από το πεδίο· στον ζυγό σύγκρισης το g απαλείφεται, στο δυναμόμετρο όχι· παράδειγμα με τον Δία
Σωστή επιλογή (2/3)	mg = kx → x = mg/k· ισχύει για ίδιο ελατήριο και ίδιο g
Αναφορά (3/2–3/4)	k ως «σκληρότητα»· λόγος βαρών = λόγος μαζών σε κάθε πλανήτη, αφού το g απαλείφεται

Ως προς τη διαλογική δομή (ΕΕ4), η ακολουθία Ζωή → Κύριος Γιώργος → Ελισάβετ → Κύριος Γιώργος τηρήθηκε σε όλες τις ερωτήσεις. Ο Κύριος Γιώργος έδωσε τη διδακτική εξήγηση, η Ζωή έκανε σωκρατικές ερωτήσεις και σε μία περίπτωση συνέδεσε τη συζήτηση με την καθημερινή γλώσσα («στην καθημερινή γλώσσα λέμε "ζυγίζω 50 kg" — αυτό δεν είναι λάθος;»), και η Ελισάβετ τον κριτικό έλεγχο, με ακραίες περιπτώσεις όπως ένα διαστημόπλοιο μακριά από κάθε πλανήτη. Στο σενάριο αναφοράς οι ερωτήσεις της Ελισάβετ ήταν επιβεβαιωτικές («Άρα…;»), σε συμφωνία με τις ορθές απαντήσεις. Όπου τέθηκε ερώτηση κατανόησης δόθηκε θετική απάντηση, οπότε ο μηχανισμός επανεξήγησης δεν ενεργοποιήθηκε.

E. Συζήτηση

Η ανάλυση έδειξε ότι ο C λειτούργησε ως διακριτή διάσταση της αξιολόγησης: κατέγραψε τη ρητή αντίφαση του Σεναρίου 2 και απέδωσε C = 1,00 στο σενάριο αναφοράς, ενώ η επικάλυψη C και V στο Σενάριο 1 υποδεικνύει τη βελτίωση των C1–C5 (Ενότητα IV.D). Οι τιμές περιγράφουν τις συγκεκριμένες εκτελέσεις και αποτελούν βάση για την εφαρμογή του συστήματος σε πραγματικό μαθητικό πληθυσμό.

IV. ΣΥΜΠΕΡΑΣΜΑΤΑ

A. Απαντήσεις στα Ερευνητικά Ερωτήματα

ΕΕ1 Το σύστημα εστίασε στη διάκριση μάζας και βάρους και στην εξάρτηση του βάρους από το βαρυτικό πεδίο. Η αιτιολόγηση έδωσε περισσότερο υλικό για εξέταση από την επιλογή μόνη της· οι προκαθορισμένες παρανοήσεις αποτυπώθηκαν με χαμηλές τιμές V και L, ενώ οι ορθές απαντήσεις του σεναρίου αναφοράς πήραν υψηλές τιμές.

ΕΕ2 Η ανατροφοδότηση περιέλαβε αποσαφήνιση φυσικών εννοιών, εξήγηση σχέσεων και παραδείγματα, χαρακτηριστικά που αντιστοιχούν στο επίπεδο έργου και διαδικασίας (Hattie & Timperley, 2007; Shute, 2008).

ΕΕ3 Κάθε δείκτης έδωσε διαφορετική όψη της αιτιολόγησης: ο V την επιστημονική ορθότητα, ο L τη λογική ακολουθία, ο R την καταλληλότητα για τη βαθμίδα, η Cl τη σαφήνεια και ο C τη συνέπεια μεταξύ των αιτιολογήσεων. Τα Σενάρια 1 και 2 είχαν το ίδιο ποσοστό ορθών επιλογών (1/5) και την ίδια τιμή C, αλλά διαφορετικές τιμές R και Cl. Ο C μπαίνει ως κοινή τιμή στο Q όλων των ερωτήσεων της συνεδρίας και κατέγραψε τη ρητή αντίφαση του Σεναρίου 2.

ΕΕ4 Ο Κύριος Γιώργος έδωσε τη διδακτική εξήγηση, η Ζωή έκανε σωκρατικές ερωτήσεις και η Ελισάβετ

τον κριτικό έλεγχο· η ακολουθία αυτή, μαζί με την απαίτηση αιτιολόγησης, δημιουργεί ευκαιρίες για επέκταση και αναστοχασμό (Alexander, 2008).

Β. Συμβολή της Εργασίας

Η εργασία υλοποίησε και εξέτασε ένα πρωτότυπο που συνδυάζει σχολικές σημειώσεις, τρεις διαλογικούς παιδαγωγικούς ρόλους, απαίτηση αιτιολόγησης και το πλαίσιο VCLRCl-Q. Στις τρεις συνθετικές εκτελέσεις οι δείκτες ξεχώρισαν τη σαφήνεια από την επιστημονική ορθότητα και διέκριναν καθαρά τα σενάρια παρανοήσεων από το σενάριο αναφοράς. Η συμβολή βρίσκεται στον σχεδιασμό αυτής της διαδικασίας και στην πρώτη, διερευνητική εφαρμογή της. Στη *διδακτική των Φυσικών Επιστημών*, η βιβλιογραφία για τις εναλλακτικές αντιλήψεις στη μάζα και στο βάρος αξιοποιήθηκε για συνθετικές εισόδους, συνδεδεμένες με την ανάλυση των αποκρίσεων του συστήματος. Στην *αξιολόγηση και ανατροφοδότηση*, η εργασία οργανώνει την ορθότητα, τη λογική ακολουθία, την καταλληλότητα, τη σαφήνεια και τη συνέπεια ως διακριτές όψεις της αιτιολόγησης· οι δείκτες αξιοποιούνται περιγραφικά, και η σχέση τους με τη μαθησιακή πρόοδο εντάσσεται σε επόμενο στάδιο. Στην *τεχνητή νοημοσύνη στην εκπαίδευση*, συνδυάζει επιλεγμένες σχολικές σημειώσεις, δομημένες οδηγίες αξιολόγησης και τρεις παιδαγωγικούς ρόλους, με το μοντέλο περιορισμένο στο υλικό του κεφαλαίου.

C. Παιδαγωγικές Προεκτάσεις

Για τον εκπαιδευτικό, η εφαρμογή μπορεί να δώσει υλικό για αιτιολογήσεις σαφείς αλλά επιστημονικά λανθασμένες, ισχυρισμούς που αντιφάσκουν μεταξύ τους, παρανοήσεις που επανέρχονται στον διάλογο και δηλώσεις άγνοιας ή αβεβαιότητας που μπορούν να γίνουν αφετηρία συζήτησης στην τάξη. Κάθε τιμή ερμηνεύεται μαζί με την ερώτηση και τις συνθήκες της, και ο εκπαιδευτικός έχει τον κεντρικό ρόλο στις παιδαγωγικές αποφάσεις. Από τον σχεδιασμό προκύπτουν αρχές που αξίζει να αξιολογηθούν περαιτέρω: σύνδεση με επιλεγμένο υλικό· αιτιολόγηση πριν από την ανατροφοδότηση, ώστε το σκεπτικό του μαθητή να γίνεται ορατό (Chin & Osborne, 2010)· ανατροφοδότηση πέρα από το σωστό/λάθος· παιδαγωγικοί ρόλοι που οργανώνουν ερωτήσεις επέκτασης και κριτικού ελέγχου.

D. Μελλοντική Έρευνα

Προτείνονται:

1. εμπειρική αξιολόγηση με μαθητές Γυμνασίου, με αρχική και τελική αξιολόγηση, ομάδα σύγκρισης και, αν είναι εφικτό, μεταγενέστερη μέτρηση της διατήρησης της γνώσης, με πρωτόκολλο ενημέρωσης, συναίνεσης και διαγραφής δεδομένων·
2. επικύρωση των δεικτών: ανεξάρτητοι ειδικοί βαθμολογούν τις ίδιες αιτιολογήσεις με την ίδια ρουμπρίκα χωρίς να βλέπουν τις τιμές του μοντέλου, με σαφέστερη οριοθέτηση των C1–C5 από τον V (π.χ. C χωρίς V και L στο consistency_log)· επαναλαμβανόμενες εκτελέσεις για τη σταθερότητα των βαθμών· σύνδεση των βαθμών με ανεξάρτητες δοκιμασίες γνώσεων·
3. επέκταση σε άλλες ενότητες (Δυνάμεις, Ηλεκτρισμός, Κύματα, Οπτική), τάξεις και μαθήματα·
4. τεχνολογικές βελτιώσεις: αποθήκευση ολόκληρου του κουίζ, υπολογισμός των μέσων όρων και του Q στον διακομιστή και πίνακας παρακολούθησης για τον εκπαιδευτικό.

Η αρχιτεκτονική της ρουμπρίκας VCLRCl-Q επιτρέπει την προσαρμογή της σε διαφορετικά γνωστικά αντικείμενα. Η μεταφορά δεν απαιτεί αλλαγή της δομής των πέντε δεικτών, των βαρών ή του τύπου του Q, αλλά προσαρμογή της διατύπωσης των επιμέρους ελέγχων (π.χ. ο V1 γίνεται «σωστός νόμος, κανόνας ή ορισμός του μαθήματος» και ο Cl3 «σωστή ορολογία του μαθήματος») και αντικατάσταση της γνωστικής βάσης με τις σημειώσεις του νέου μαθήματος. Έτσι ο μηχανισμός αξιολόγησης μένει ενιαίος, ενώ τα κριτήρια εναρμονίζονται με τις απαιτήσεις κάθε αντικειμένου. Η αρχιτεκτονική αυτή υποστηρίζει την τεχνική επεκτασιμότητα του συστήματος· η αξιοπιστία της αξιολόγησης σε κάθε νέο μάθημα αποτελεί αντικείμενο ξεχωριστής εμπειρικής διερεύνησης.

Τα επόμενα βήματα είναι η αξιολόγηση των δεικτών από ανεξάρτητους ειδικούς και η εφαρμογή με πραγματικούς μαθητές, ώστε να εξεταστεί η συμβολή του συστήματος στη μάθηση.

ΑΝΑΦΟΡΕΣ

Alexander, R. J. (2008). *Towards dialogic teaching: Rethinking classroom talk* (4th ed.). Dialogos.

Chin, C., & Osborne, J. (2010). Supporting argumentation through students' questions: Case studies in science classrooms. *Journal of the Learning Sciences, 19*(2), 230–284. https://doi.org/10.1080/10508400903530036

Driver, R., Guesne, E., & Tiberghien, A. (Eds.). (1985). *Children's ideas in science*. Open University Press.

Galili, I. (1993). Weight and gravity: Teachers' ambiguity and students' confusion about the concepts. *International Journal of Science Education, 15*(2), 149–162. https://doi.org/10.1080/0950069930150203

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612–618. https://doi.org/10.1109/TE.2005.856149

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211–227. https://doi.org/10.1002/sce.3730660207

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Viennot, L. (1979). Spontaneous reasoning in elementary dynamics. *European Journal of Science Education, 1*(2), 205–221. https://doi.org/10.1080/0140528790010209

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems, 36*.