COMPETIȚIA DE PROIECTE DE CERCETARE A ACADEMIEI OAMENILOR DE ȘTIINȚĂ DIN ROMÂNIA DESTINATĂ TINERILOR CERCETĂTORI „AOȘR-TEAMS-IV” EDIȚIA 2025-2026 „TRANSFORMAREA DIGITALĂ ÎN ȘTIINȚE”

Platformă web pentru generarea automată de întrebări (QuizTools)

Raport 1

Director de proiect: Conf. dr. ing. Ștefan Rușeți
Membru: Sl. dr. ing. Răzvan Păroiu
Membru: As. drd. ing. Andreea Duțulescu

Cuprins

1 Introducere

Generarea automată de întrebări prezintă un potențial ridicat pentru optimizarea proceselor de evaluare educațională, furnizarea de feedback imediat și consolidarea înțelegerii conținutului de învățare. Totuși, dezvoltarea unor întrebări de calitate, bine aliniate la obiectivele pedagogice, rămâne o provocare semnificativă. În acest context, prezentul raport descrie procesul dezvoltării unei metode pentru generarea de întrebări cu grilă multiplă (MCQ), axată pe valorificarea progreselor recente în domeniul modelelor mari de limbaj (LLM). Accentul este pus pe creșterea calității întrebărilor generate prin îmbunătățirea opțiunilor de răspuns (răspunsul corect și distractorii aferenți).

Deși metodele moderne au dus la îmbunătățiri semnificative în ceea ce privește formularea lingvistică și claritatea întrebărilor de tip MCQ, generarea automată a variantelor de răspuns din punct de vedere educațional continuă să fie o sarcină dificilă. Întrebările de calitate trebuie să evalueze acuratețea conceptuală a cunoștințelor, oferind în același timp oportunități pentru reflecție critică. Distractorii trebuie proiectați astfel încât să reflecte erori frecvent întâlnite în procesul de învățare, evitând în același timp ambiguitățile. În plus, utilizarea predominantă a modelelor LLM comerciale limitează accesibilitatea tehnologiei pentru instituțiile cu resurse restrânse. Diversitatea domeniilor educaționale implică cerințe variate asupra sistemelor de generare, care nu sunt în prezent suficient acoperite de metodele existente. Acest context evidențiază nevoia unor soluții open-source scalabile și ușor de integrat.

O analiză recentă realizată de Alhazmi et al. (2024) a identificat deficiențe persistente în generarea de întrebări MCQ, în special în ceea ce privește calitatea distractorilor. Tehnicile bazate pe modele lingvistice pre-antrenate (Gao et al., 2019; Shuai et al., 2023; Maurya & Desarkar, 2020), chiar și atunci când sunt adaptate pe seturi de date specializate, produc adesea distractori redundanți sau lipsiți de diversitate semantică. Alte abordări (Wang et al., 2023; Jiang & Lee, 2017) prezintă riscul de a introduce elemente din răspunsul corect în distractori, ceea ce poate compromite validitatea întrebării.

În acest cadru, propunem o soluție nouă pentru generarea automată a întrebărilor MCQ, orientată spre eliminarea acestor limitări. Metodologia se bazează pe integrarea principiilor pedagogice în generarea automată, cu accent pe formularea clară a raționamentului asociat răspunsurilor corecte și proiectarea unor distractori care reflectă greșeli cognitive frecvente în rândul elevilor. Soluția utilizează modele open-source, cărora li s-a aplicat un proces de fine-tuning pe seturi de date publice, augmentate cu explicații generate sintetic. Acest proces de augmentare vizează îmbunătățirea capacității modelului de a produce întrebări relevante din punct de vedere educațional. În plus, arhitectura propusă este optimizată pentru eficiență computațională,

permițând rularea pe infrastructuri cu cost redus, fără dependență de servicii comerciale. Procesul de generare include și un mecanism de filtrare în mai multe etape pentru asigurarea calității întrebărilor și clasificarea rezultatelor în funcție de nivelul de încredere estimat de model. Prin tratarea sistematică a limitărilor identificate în lucrările anterioare, abordarea noastră urmărește atât îmbunătățirea performanței sistemelor de generare, cât și creșterea aplicabilității lor practice. În sprijinul reproductibilității, sunt puse la dispoziție în regim open-source toate modelele, codul asociat procesului de antrenare, precum și un set de date extins, cuprinzând peste 300.000 de întrebări augmentate cu explicații și erori cognitive sintetice [1].

2 Stadiul curent al cercetării

2.1 Generarea de întrebări pentru contexte educațional

Generarea automată de întrebări a fost intens studiată în scopuri educaționale, cu accent pe îmbunătățirea scalabilității și calității evaluărilor automate. Lucrarea realizată de Bulathwela et al. (2023) propune o strategie bazată pe utilizarea modelului T5-small, adaptat prin pre-antrenare continuă pe corpusuri științifice, pentru a optimiza capacitatea modelului de a genera întrebări în limbaj academic. Abordarea se distinge de metodele tradiționale prin utilizarea de seturi de date specializate, nu generale, ceea ce a condus la o performanță crescută în generarea de întrebări educaționale relevante.

Într-o altă direcție, Hwang et al. (2023) analizează generarea întrebărilor cu răspunsuri multiple aliniate cu nivelele taxonomiei Bloom, utilizând GPT-3.5 cu prompting de tip zero-shot. Validarea s-a realizat printr-o combinație de clasificatori automați (RoBERTa), verificare de conformitate bazată pe reguli și evaluare umană, pentru a asigura corectitudinea taxonomică și calitatea formulării.

Scaria et al. (2024) propun o evaluare comparativă între mai multe modele, inclusiv GPT-4, PaLM 2, LLaMA2 și Mistral, pentru generarea de întrebări corespunzătoare diferitelor niveluri cognitive. Tehnicile de prompting au inclus explicații explicite, raționamente intermediare (Chain-of-Thought) și exemple umane. Calitatea rezultatelor a fost analizată automat prin Gemini Pro, concentrându-se pe validitate și acoperirea nivelurilor cognitive.

O altă direcție investigată de Cui et al. (2024) a vizat inserarea întrebărilor direct în textul educațional, cu scopul de a stimula atenția și înțelegerea utilizatorilor. Setul de date GUIDINGQ, format din întrebări extrase din materiale educaționale, a fost folosit pentru antrenarea modelului Flan-T5. Metoda propusă accentuează integrarea contextuală a întrebărilor, în contrast cu generarea izolată, contribuind la coeziunea dintre conținut și întrebări.

În domeniul întrebărilor de matematică, Fernandez et al. (2024) au

[1] `https://github.com/upb-nlp/AIED-MCQ-with-Explanations`

dezvoltat DiVERT, un sistem care modelează explicit erorile elevilor pentru a genera distractori relevanți. Acesta include trei componente specializate: una pentru predicția erorilor, una pentru generarea distractorilor și una pentru evaluarea calității acestora. DiVERT se diferențiază de metodele anterioare prin abordarea sa centrată pe cauzalitatea greșelilor, ceea ce sporește interpretabilitatea și relevanța educațională. În comparațiile experimentale, DiVERT a depășit performanța GPT-4o în generarea de distractori similari celor proiectați de experți umani.

În ciuda progreselor notabile, rămân limitări importante în ceea ce privește utilizarea extensivă a modelelor comerciale, precum și dependența de metode mai vechi sau nescalabile. Aceste constrângeri subliniază nevoia dezvoltării unor soluții open-source robuste, cu costuri reduse și aplicabilitate extinsă în contexte educaționale variate.

2.2 Criterii de evaluare pentru generarea de întrebări

În ceea ce privește evaluarea sistemelor de generare de întrebări, studiile recente propun metrici structurate pentru a analiza eficiența modelelor. Chen et al. (2024) au introdus un set de rubrici denumit Dr. Academy, care evaluează performanțele LLM-urilor în generarea de întrebări educaționale pe trei categorii: sarcini generale, monodisciplinare și interdisciplinare. Sistemul de evaluare utilizează patru criterii: consistență, relevanță, acoperire și adecvare, iar scorurile au fost generate cu ajutorul GPT-4 pentru automatizarea procesului. Pe o direcție complementară, Fu et al. (2024) au propus o schemă de evaluare pe șapte dimensiuni, incluzând claritatea, concizia, adecvarea, coerența și consistența răspunsurilor. Modelele testate (inclusiv GPT-4, T5, BART) au fost analizate în contexte diferite, incluzând fine-tuning, LoRA și prompting zero-shot. Rezultatele au indicat performanțe ridicate pentru GPT-4, dar au evidențiat riscuri comune legate de capacitatea de răspuns și consistența logică a întrebărilor generate.

Aceste instrumente de evaluare oferă o bază solidă pentru diagnosticarea limitărilor modelelor actuale, dar relevă și dominanța netă a soluțiilor comerciale în performanță. Acest dezechilibru evidențiază importanța dezvoltării unor alternative open-source competitive, pentru a asigura democratizarea accesului la tehnologii educaționale avansate. Concluziile formulate în aceste studii sunt în acord cu rezultatele experimentale prezentate în secțiunile următoare.

3 Metodă

Calitatea întrebărilor într-un test de înțelegere a textului depinde de mai mulți parametri. În primul rând, formularea trebuie să fie clară și să permită evaluarea directă a informațiilor extrase din textul furnizat. Răspunsul asociat trebuie să reflecte o acoperire adecvată a conținutului, iar combinația

între întrebare și răspuns trebuie să contribuie semnificativ la procesul de învățare, oferind informații coerente și relevante. De asemenea, distractorii trebuie proiectați astfel încât să fie plauzibili și să nu poată fi eliminați cu ușurință pe baza unor indicii evidente. Este esențială o înțelegere clară a fiecărei componente a testului.

Pornind de la aceste cerințe, propunem un sistem de generare automată de întrebări cu variante multiple de răspuns, bazat pe: (1) formularea unei întrebări relevante în raport cu un context dat, (2) generarea răspunsului corect condiționată de un raționament explicit și justificat și (3) elaborarea distractorilor pornind de la erori de interpretare frecvente în rândul elevilor. Abordarea noastră implică antrenarea unui model de limbaj pentru a produce secvențial următoarele elemente: întrebarea propriu-zisă, raționamentul care susține răspunsul corect, răspunsul corect și ipoteze eronate folosite ca bază pentru generarea variantelor greșite.

3.1 Seturi de date și augmentare

Datele utilizate pentru antrenarea modelului de generare au fost extrase din surse publice, accesibile mediului academic. Alegerea acestora a fost datorată varietății conținutului textual și diversității domeniilor abordate, cu scopul de a obține o acoperire cât mai extinsă și adaptabilă la multiple contexte. Mai exact, au fost utilizate subseturi de antrenare provenind din următoarele seturi de date:

- **SQuAD** (Rajpurkar et al., 2016): Un corpus pentru sarcini de tip întrebare-răspuns, construit pe baza articolelor Wikipedia, cu întrebări formulate manual de către colaboratori umani.
- **HotpotQA** (Yang et al., 2018): Similar cu SQuAD prin utilizarea sursei Wikipedia, dar pune accentul pe raționamente de tip multi-hop necesare pentru obținerea răspunsului corect.
- **NarrativeQA** (Kočiskỳ et al., 2018): Un corpus orientat spre înțelegerea narativă profundă, punând accent pe entități, relații și evenimente, având ca suport rezumate umane ale unor cărți și scenarii de film.
- **FairytaleQA** (Xu et al., 2022): Conceput cu scop educațional, conține povești fictive și întrebări elaborate manual de experți în pedagogie, centrate pe aspecte narative precum relațiile cauzale și evoluția personajelor.
- **MCTest** (Richardson et al., 2013): Un corpus destinat evaluării înțelegerii automate a textului, conținând povestiri scurte și întrebări cu răspunsuri multiple, construite colaborativ.

- **RACE** (Lai et al., 2017): Set de date extras din examene standardizate de limbă engleză, în care întrebările cu răspunsuri multiple implică raționamente complexe și niveluri avansate de înțelegere.
- **EduQG** (Hadifar et al., 2023): Corpus din domeniul educațional format din întrebări cu variante multiple, construit pe baza conținutului din manualele universitare OpenStax, caracterizat prin calitate înaltă și un grad sporit de dificultate.

Din cele enumerate mai sus, corpusurile SQuAD, HotpotQA, NarrativeQA și FairytaleQA includ doar contextul, întrebarea și răspunsul corect. În schimb, corpusurile MCTest, RACE și EduQG, conțin și distractori alături de răspunsul corect. Toate corpusurile (unele cu variații minore) au fost utilizate pentru antrenare.

Ulterior, aceste corpusuri au fost extinse cu explicații justificate pentru răspunsurile corecte, cât și cu raționamentele eronate specifice fiecărui distractor. Aceste explicații au fost generate automat de un model neuronal Llama 3.1 70B Dubey et al. 2024 cu cuantizare la 4 biți (Jacob et al., 2018).

În cadrul corpusului RACE, analiza empirică a identificat o abundență de întrebări cu structură repetitivă. Formulări precum *Care este ideea principală a pasajului?* sau *Pasajul este cel mai probabil extras din ___.* apar cu o frecvență ridicată, generând astfel o distribuție dezechilibrată a tipologiilor de întrebări. Pentru a reduce impactul acestui dezechilibru asupra procesului de antrenare, s-a decis includerea doar a unui procent restrâns din aceste instanțe, situat între 5% și 10% din totalul inițial. În plus, pentru a respecta limitele de procesare impuse de infrastructura hardware utilizată, au fost eliminate toate textele care depășeau pragurile tehnice admise. Structura finală a setului de date este prezentată în Tabelul 1.

Tabelul 1. Dimensiunile seturilor de date.

Setul de date	**Dim. antrenare**	**Dim. validare**
SQuaD	87 391	10 544
HotpotQA	89 776	7 356
NarrativeQA	65 362	3 452
FairytaleQA	8 548	1 025
MCTest	1 200	200
CURSĂ	82 173	4 587
EduQG	2 725	0
Total	337 175	27 164

3.2 Antrenare

Am antrenat un model de limbaj de tip LLM, open-source, pentru a genera simultan următoarele componente, pornind de la un context oferit ca date

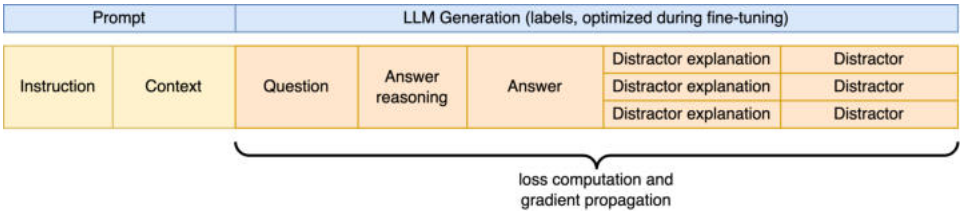

Figura 1. Formatul promptului de antrenare.

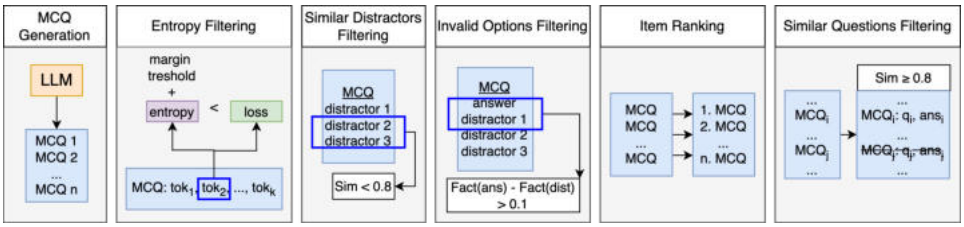

Figura 2. Sistem de generare a întrebărilor.

de intrare: o întrebare, raționamentul aferent, răspunsul corect, posibile concepții greșite și distractori relevanți. Procesul de generare urmează un format general de prompt, ilustrat în Figura 1.

Pentru desfășurarea experimentelor, am folosit Llama 3.1 8B - Instruct[2]. Antrenarea a fost realizată pe parcursul unei singure epoci, în regim de precizie mixtă (`bf16`), în conformitate cu practicile standard pentru modelele Llama. Am folosit o rată de învățare constantă, de 10^{-6}, optimizatorul AdamW, în varianta pe 8 biți și un batch size efectiv de 64, obținut prin acumulare de gradienți.

3.3 Inferență și filtrare

Am conceput un proces etapizat pentru a genera întrebări de tip grilă pornind de la un anumit context, folosind un model preantrenat (a se vedea Figura 2). Metodele de generare și selecție a întrebărilor sunt detaliate în secțiunile următoare.

Generare eșantion. Pentru fiecare context furnizat, au fost generate aproximativ 60 de întrebări. Fiecare item conține formularea întrebării, justificarea răspunsului corect, răspunsul propriu-zis, precum și o descriere a concepțiilor greșite asociate distractorilor, alături de distractorii înșiși. Am aplicat strategia de decodare **min-p** (Nguyen et al., 2024), care implică o trunchiere dinamică adaptivă, calibrată în funcție de încrederea modelului. Această abordare ajustează pragul de selecție în funcție de probabilitatea token-ului cel mai plauzibil, asigurându-se că raportul dintre probabilitatea token-ului generat și a celui mai probabil este mai mare decât *min-p*. Astfel,

[2] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

se evită includerea cuvintelor improbabile, susceptibile de a produce erori.

Filtrare bazată pe entropie. O metodă simplă de a evalua încrederea modelului în privința unei secvențe generate constă în calcularea sumei probabilităților logaritmice asociate fiecărui token. Totuși, dacă modelul produce un token greșit, este adesea predispus să continue generarea în aceeași direcție, menținând totodată un nivel ridicat de încredere pentru tokenii următori. În consecință, evaluarea probabilității totale a secvenței poate fi eronată și nu reflectă cu acuratețe validitatea sau corectitudinea conținutului generat.

În consecință, propunem examinarea probabilităților individuale ale token-ilor generați pentru a identifica posibile erori. Primul mecanism de filtrare este reprezentat de pragul *min-p* aplicat direct în timpul inferenței. Al doilea mecanism se bazează pe entropia unui token definită ca $-\sum_{t \in V} p(t) log(p(t))$ și care poate fi interpretată ca valoarea așteptată a probabilității logaritmice negative pentru un anumit token. Ulterior, comparăm probabilitățile logaritmice negative ale token-ilor generați cu entropia corespunzătoare, eliminând secvențele ce conțin tokeni semnificativ mai puțin probabili decât ar sugera valoarea așteptată

Filtrarea întrebărilor cu distractorilor similari. Pentru a evita prezența distractorilor duplicat în cadrul aceleiași întrebări, am filtrat acele întrebări care conțineau distractori cu un grad ridicat de similitudine între ei. Măsurarea similitudinii dintre doi distractori s-a realizat prin obținerea reprezentărilor latente utilizând Sentence Transformers[3] urmată de calculul similarității cosinus. Întrebările în care cel puțin două variante de răspuns prezentau o similitudine peste un prag stabilit au fost eliminate din setul final.

Filtrarea întrebărilor cu opțiunilor invalide. În cazul întrebărilor cu alegere multiplă, este esențial ca doar o singură variantă să fie corectă, iar distractorii să nu poată fi interpretați ca răspunsuri valide. Prezența mai multor opțiuni corecte poate genera ambiguitate și confuzie în rândul studenților. Pentru a identifica și elimina astfel de situații, am verificat ca scorul de factualitate al fiecărui distractor să fie mai mic decât cel al răspunsului corect desemnat. În acest scop, am utilizat **MiniCheck**[4] (Tang et al., 2024), un model compact de verificare a veridicității, pentru a calcula scorul de factualitate al fiecărei variante de răspuns în raport cu dovezile extrase din context și întrebare.

Ordonarea întrebărilor. Pentru a garanta calitatea ridicată a întrebărilor generate, am ordonat întrebările rămase astfel încât cele mai relevante să fie returnate printre primele atunci când un utilizator solicită un set limitat de întrebări. Clasificarea se realizează pe baza probabilității logaritmice a conținutului generat, incluzând întrebări, răspunsuri, distractorii și

[3] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

[4] `https://huggingface.co/bespokelabs/Bespoke-MiniCheck-7B`

explicațiile aferente.

În majoritatea cazurilor, întrebările sunt ordonate descrescător după $log(P_{model}(item \mid prompt + context))$, unde un item corespunde secvenței complete generate, incluzând întrebări, răspunsuri, distractorii și explicațiile aferente.

Filtrarea întrebărilor similare. Ca ultim pas al procesului, eliminăm întrebările duplicate. Similar abordării folosite pentru filtrarea distractorilor redundanți, calculăm similaritatea cosinus între reprezentările latente ale întrebărilor și ale răspunsurilor asociate. În cazul în care sunt identificate întrebări duplicate, păstrăm doar itemul cu scorul global superior, iar celelalte sunt eliminate din setul final.

4 Evaluarea performanței

Pentru a evalua performanța metodei propuse, am utilizat un set de date din Sistemul Inteligent de Meditații iSTART (Perret et al., 2017) destinat învățământului preuniversitar (K-12). Analiza s-a bazat pe un subset cu 55 de texte care acoperă o varietate de domenii, inclusiv istorie, geografie, știință și tehnologie. Fiecare text are, în medie, aproximativ 500 de cuvinte, corespunzând unui număr de circa 30 de propoziții.

4.1 Baselines

În vederea evaluării calității metodei noastre de generare a întrebărilor cu răspunsuri multiple (MCQ), am comparat rezultatele obținute din trei metode alternative. Pentru fiecare metodă, au fost generate întrebări pe baza acelorași contexte, ceea ce ne-a permis să analizăm performanța modelului nostru în raport atât cu soluții automatizate existente, cât și cu întrebări redactate de experți umani.

Fără explicații. Am investigat influența integrării raționamentului și explicațiilor în timpul antrenamentului și inferenței asupra calității întrebărilor generate. În acest scop, am realizat un studiu de ablație, antrenând modelul pentru a genera exclusiv întrebările, răspunsurile corecte și distractorii; astfel, am exclus justificarea răspunsului corect și explicațiile privind concepțiile greșite asociate distractorilor. Antrenarea acestui model a urmat același format și a utilizat aceiași hiperparametri ca la modelul propus, diferența constând în excluderea componentelor de raționament și explicație din prompt.

GPT-4o. În calitate de standard industrial, am utilizat GPT-4o, un model proprietar cu costuri asociate procesului de inferență. Modelului i s-a furnizat aceeași structură de intrare ca în cazul metodei noastre, fiind instruit să genereze un set complet de întrebări cu răspunsuri multiple. Acest set a inclus formularea întrebării, raționamentul aferent răspunsului corect,

răspunsul corect propriu-zis, concepții greșite frecvente și distractorii corespunzători.

Întrebări umane. Setul de date a conținut întrebări cu răspunsuri multiple utilizate ca etalon pentru evaluarea calitativă a întrebărilor generate automat.

4.2 Configurarea evaluării

Pentru a evalua calitatea întrebărilor cu răspunsuri multiple generate prin diverse metode (inclusiv modelul nostru cu explicații, o variantă simplificată a aceluiași model fără explicații, GPT-4o și întrebări cu răspunsuri multiple redactate de experți umani), am adoptat o abordare de comparație pe perechi. În cadrul acestui proces, anotatori umani au evaluat perechi de întrebări din surse diferite. Comparația pe perechi reprezintă o metodă de evaluare larg utilizată, în special în sarcini care implică judecăți subiective de calitate, fiind aplicată frecvent în evaluarea sistemelor automate. Studii anterioare au demonstrat eficiența acestei tehnici în contexte bazate pe preferințe, în care experții joacă rolul de arbitri ai calității (Qin et al., 2024; Liusie et al., 2024).

Spre deosebire de scalele de evaluare absolute, comparația pe perechi oferă rezultate mai explicite, indicând în mod direct care dintre metode generează itemi de calitate superioară. Totodată, această abordare permite existența unor egalități atunci când întrebările, deși diferite tematic, prezintă un nivel similar de calitate. Un avantaj al acestei metode este reducerea efortului cognitiv al adnotatorilor, deoarece compararea directă a doi itemi este mai intuitivă și mai consistentă decât atribuirea unor scoruri individuale fiecărui item în parte.

Evaluarea umană a fost realizată cu ajutorul a 5 anotatori, care au analizat calitatea întrebărilor cu răspunsuri multiple generate prin diferite metode. Toți anotatorii erau vorbitori non-nativi de limba engleză și aveau experiență anterioară în activități educaționale desfășurate cu studenții. Fiecărui anotator i-au fost alocate 15 texte, pentru care au efectuat câte 30 de comparații pereche per text.

În etapa inițială, toți anotatorii au evaluat un set de cinci texte (n = 150 de perechi de comparații) cu scopul de a verifica înțelegerea corectă a sarcinii de către evaluatori. Concordanța inter-evaluatori a fost măsurată utilizând atât procentul de concordanță cât și Kappa lui Conger (Conger, 1980), o extensie a măsurii Kappa lui Cohen pentru evaluatori multipli. Concordanța a fost calculată în două variante: cu luarea în considerare a răspunsurilor de tip egalitate, cât și fără acestea. Procentul de acord (n = 150 de perechi) a fost de 49%, iar măsura Kappa lui Conger a înregistrat o valoare de 0,19 [CI 95% 0,14 - 0,24], considerată o concordanță ușoară conform interpretării lui Landis și Koch (Landis & Koch, 1977). Procentul de concordanță în varianta fără egalitate (n = 65 de perechi) a crescut la 64%, cu un Ka-

ppa lui Conger de 0,29 [CI 95% 0,17-0,41], considerată drept concordanță moderată (Landis & Koch, 1977). Astfel de acorduri sunt foarte frecvente în evaluarea umană a textelor generate automat, conform observațiilor lui Amidei et al. (2018). În plus, am urmărit îmbunătățirea validității prin desfășurarea comparațiilor într-un cadru cât mai apropiat de situații reale, în care ajustările legate de subiectivitate nu sunt uzuale. Astfel, faza de calibrare a avut ca scop principal asigurarea înțelegerii clare a instrucțiunilor de către fiecare evaluator, fără o uniformizare strictă a răspunsurilor între evaluatori. Această abordare le-a permis evaluatorilor să își păstreze interpretările individuale, menținând în același timp un obiectiv comun, dar reflectând mai fidel condițiile practice de utilizare. După finalizarea etapei de calibrare, fiecare adnotator a continuat în mod independent evaluarea, adnotând câte 10 texte distincte.

Pentru fiecare text, au fost realizate 30 de comparații perechi, generate conform următoarei proceduri: fiecare dintre cele patru metode (modelul nostru cu explicații, varianta fără explicații, GPT-4o și întrebările cu răspunsuri multiple redactare manual de oameni) a furnizat câte 5 întrebări per text. Aceste întrebări au fost apoi asociate sistematic, astfel încât fiecare comparație să includă întrebări provenind din metode diferite. Mai precis, fiecare întrebare generată de o anumită metodă a fost asociată cu câte o întrebare provenind din celelalte trei metode, asigurându-se o distribuție echilibrată a comparațiilor și maximizând evaluarea între metode.

În etapa de adnotare, evaluatorii au primit instrucțiuni să aleagă întrebarea cu răspunsuri multiple care se evidenția prin calitate superioară sau, dacă nu era posibilă o diferențiere clară, să opteze pentru varianta de egalitate. Criteriile care au stat la baza deciziilor au vizat patru aspecte: corectitudinea unică a răspunsului notat întotdeauna primul, coerența întrebării și a opțiunilor cu informațiile din textul-sursă, nivelul de dificultate non-trivial al întrebării, precum și capacitatea acesteia de a testa înțelegerea, nu simpla reamintire a informațiilor. Evaluatorii au fost îndrumați să selecteze itemul care respecta cel mai mare număr dintre aceste criterii, iar în situații în care diferențierea era imposibilă sau extrem de dificilă, să recurgă la marcarea egalității. Nu li s-a cerut să indice explicit care criterii erau îndeplinite de fiecare întrebare, pentru a menține caracterul natural al sarcinii, similar unui context aplicativ real, în care utilizatorii nu sunt obligați să justifice alegerile sub presiunea sarcinii cognitive.

Ulterior, aceste comparații în perechi au fost analizate utilizând modelul probabilistic Bradley-Terry (Bradley & Terry, 1952), cu scopul de a obține o ierarhizare și scoruri asociate fiecărui set de întrebări. Modelul permite estimarea forței relative a unor elemente pe baza probabilităților rezultate din comparații binare între acestea.

5 Rezultate

Figura 3 oferă o prezentare detaliată a rezultatelor procesului de evaluare, ilustrând performanța relativă a fiecărei metode în generarea de întrebări. Diagrama reflectă rezultatele obținute în urma comparațiilor pereche dintre întrebările produse de două metode distincte. Fiecare comparație a fost clasificată drept „Câștig" în cazul în care întrebarea generată de prima metodă a fost aleasă, respectiv „Pierdere" atunci când a fost selectată întrebarea celei de-a doua metode. Pentru fiecare pereche de metode s-au realizat 275 de comparații, iar barele indică atât valorile absolute, cât și procentele aferente câștigurilor, înfrângerilor și egalităților în cadrul acestor comparații.

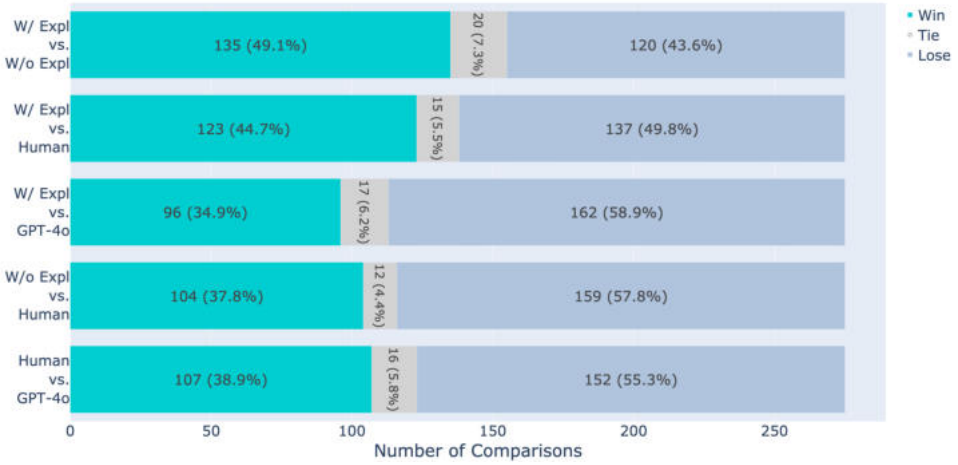

Figura 3. Rezultate comparații perechi.

Pe baza modelului Bradley-Terry, s-a obținut un clasament al metodelor analizate, exprimat prin scoruri care reflectă performanța relativă: GPT-4o (2,89), întrebările redactate manual de evaluatori umani (0,86), varianta propusă cu explicații (0,70) și varianta propusă fără explicații (0,57). Aceste valori cuantifică probabilitatea ca o metodă să fie preferată în comparație cu celelalte, în funcție de datele colectate din evaluările pereche.

6 Discuție

Analiza bazată pe Figura 3 arată că varianta propusă care integrează explicații depășește ablația (fără explicații), susținând ipoteza conform căreia includerea raționamentului asociat răspunsului corect și distractorilor în prompturile de antrenare și în etapa de inferență are un impact pozitiv asupra calității itemilor generați. Această abordare contribuie la formularea unor întrebări mai riguroase, în care opțiunile de răspuns sunt mai bine motivate, iar distractorii lipsiți de validitate sunt reduși. Creșterea calității este

atribuită utilizării explicațiilor sintetice în etapa de adnotare a datelor de instruire, ceea ce demonstrează potențialul generării automate a explicațiilor în optimizarea generării MCQ-urilor, fără a necesita un efort suplimentar semnificativ din partea adnotatorilor umani.

În același timp, rezultatele obținute confirmă că modelul GPT-4o obține cele mai bune scoruri la nivel global. Această performanță este explicabilă prin dimensiunea superioară a modelului și prin procesul extins de aliniere la preferințele utilizatorilor umani. Prin contrast, soluția propusă este construită pe baza unui model open-source de dimensiuni reduse. Obținerea unor rezultate comparabile cu acest model mai compact este relevantă în contextul democratizării accesului la instrumente educaționale eficiente și ieftine, reducând astfel dependența de sisteme proprietare. În plus, utilizarea unui model cu o amprentă computațională redusă susține obiectivele de eficiență energetică și sustenabilitate, având implicații favorabile în contextul scalabilității practice.

Întrebările redactate manual de evaluatori umani înregistrează o performanță ușor superioară metodei noastre cu explicații, ceea ce evidențiază dificultatea replicării prin modele automate a elementelor de expertiză, rafinament logic și structurare conceptuală care caracterizează itemii creați de experți. Diferența devine mai accentuată atunci când se compară întrebările umane cu cele generate fără explicații, susținând încă o dată beneficiile includerii explicațiilor în procesul automat de generare.

Limitări: În ceea ce privește limitările, trebuie menționat că, deși integrarea explicațiilor contribuie la rafinarea distractorilor, sunt necesare investigații suplimentare pentru a asigura că aceștia reflectă fidel erorile conceptuale frecvente ale elevilor din diverse domenii și niveluri educaționale. În plus, sistemele de filtrare și clasificare implementate pot exclude unele întrebări potențial relevante, ca urmare a setărilor de prag utilizate. O altă limitare derivă din dimensiunea redusă a grupului de evaluatori și din absența vorbitorilor nativi de limba engleză, ceea ce afectează validitatea externă a evaluării. Cercetările viitoare pot investiga mecanisme adaptive de stabilire a pragurilor sau strategii de învățare prin consolidare pentru îmbunătățirea selecției itemilor. În plus, deși acest studiu s-a bazat pe seturi de date deja existente, dezvoltarea unor corpusuri extinse cu explicații adnotate ar putea susține o antrenare mai eficientă și performanțe superioare în generarea MCQ-urilor.

7 Concluzii și activități viitoare

Abordarea propusă demonstrează că este posibilă generarea de întrebări cu răspuns multiplu de înaltă calitate prin antrenarea unui model lingvistic de mari dimensiuni (LLM) open-source, utilizând seturi de date pu-

blice completate cu explicații sintetice de raționament. Această metodă sporește calitatea MCQ-urilor prin integrarea explicită a justificărilor pentru răspunsurile corecte și a distractorilor concepuți pe baza concepțiilor greșite frecvent întâlnite. În acest mod, se abordează provocări specifice ale domeniului, precum necesitatea de a furniza studenților explicații care evidențiază de ce anumite opțiuni sunt corecte și de a reduce generarea de distractori nerelevanți, prin constrângerea modelului să țină cont de erori de raționament frecvente. În plus, un proces de filtrare în mai multe etape contribuie la asigurarea validității și soliditații întrebărilor finale. Această strategie sprijină nu doar validitatea întrebărilor generate, ci și dezvoltarea unei înțelegeri conceptuale mai profunde a textului de către elevi.

Un beneficiu suplimentar al utilizării concepțiilor greșite drept cadru pentru generarea de distractori constă în creșterea diversității acestora, precum și în posibilitatea de a controla caracteristicile acestora în funcție de obiectivele pedagogice. În perspectivă, se intenționează extinderea controlului acordat utilizatorilor prin introducerea unor metadate suplimentare, cum ar fi nivelul de dificultate al întrebărilor sau nivelul educațional țintă.

Rezultatele evaluărilor umane susțin eficacitatea metodei propuse, indicând că modelul antrenat depășește metodele tradiționale de antrenare și produce MCQ-uri cu o valoare educațională mai mare. Deși performanța acestuia rămâne ușor inferioară întrebărilor redactate de oameni, diferența s-a redus considerabil, iar metoda oferă o soluție automatizată scalabilă pentru generarea de evaluări educaționale. Având la bază un model open-source, soluția propusă nu depinde de infrastructură proprietară, ceea ce o face mai accesibilă și mai ușor de adaptat în diferite contexte educaționale.

O potențială direcție de cercetare viitoare este integrarea optimizării directe a preferințelor (DPO) (Rafailov et al., 2024) în procesul de generare. Dat fiind că evaluarea calității itemilor a fost realizată pe baza comparațiilor pereche, aceste date pot servi ca fundament pentru antrenarea suplimentară a modelului, aliniind generarea de întrebări la preferințele exprimate implicit de evaluatori umani. Aplicarea DPO ar putea îmbunătăți semnificativ relevanța și calitatea MCQ-urilor generate, reflectând mai fidel cerințele reale ale utilizatorilor. Totuși, un obstacol major în această direcție este necesitatea existenței unui volum mare de date cu comparații în perechi, care ar putea fi colectate prin strategii precum LLM-as-a-judge (utilizarea unui LLM ca evaluator automat).

În plus, o evaluare extinsă, care să includă analiza calității întrebărilor generate împreună cu raționamentele asociate, ar aduce o înțelegere mai profundă a valorii pedagogice a metodei. Implicarea directă a studenților în acest proces ar permite măsurarea impactului explicațiilor asupra înțelegerii și retenției, completând evaluările actuale și orientând îmbunătățirile ulterioare.

Bibliografie

Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. (2024, November). Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 14437–14458). Miami, Florida, USA: Association for Computational Linguistics.

Amidei, J., Piwek, P., & Willis, A. (2018). Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th international conference on computational linguistics* (pp. 3318–3329).

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Bulathwela, S., Muse, H., & Yilmaz, E. (2023). Scalable educational question generation with pre-trained language models. In *International conference on artificial intelligence in education* (pp. 327–339).

Chen, Y., Wu, C., Yan, S., Liu, P., & Xiao, Y. (2024). Dr. academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 3138–3167).

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological bulletin*, *88*(2), 322.

Cui, P., Zouhar, V., Zhang, X., & Sachan, M. (2024). How to engage your readers? generating guiding questions to promote active reading. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 11749–11765).

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... others (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fernandez, N., Scarlatos, A., Feng, W., Woodhead, S., & Lan, A. (2024). Divert: Distractor generation with variational errors represented as text for math multiple-choice questions. In *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 9063–9081).

Fu, W., Wei, B., Hu, J., Cai, Z., & Liu, J. (2024, November). QGEval: Benchmarking multi-dimensional evaluation for question generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp.

11783–11803). Miami, Florida, USA: Association for Computational Linguistics.

Gao, Y., Bing, L., Li, P., King, I., & Lyu, M. R. (2019). Generating distractors for reading comprehension questions from real examinations. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 6423–6430).

Hadifar, A., Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023). Eduqg: A multi-format multiple-choice dataset for the educational domain. *IEEE Access*, *11*, 20885–20896.

Hwang, K., Challagundla, S., Chen, L. K., & Choa, F.-S. (2023). Towards ai-assisted multiple choice question generation and quality evaluation at scale: Aligning with bloom's taxonomy. In *Workshop on generative ai for education.*

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *2018 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (pp. 2704–2713).

Jiang, S., & Lee, J. S. (2017). Distractor generation for chinese fill-in-the-blank items. In *Proceedings of the 12th workshop on innovative use of nlp for building educational applications* (pp. 143–148).

Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., & Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, *6*, 317–328.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 785–794).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Liusie, A., Manakul, P., & Gales, M. (2024). Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)* (pp. 139–151).

Maurya, K. K., & Desarkar, M. S. (2020). Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors

for multiple-choice questions for reading comprehension. In *Proceedings of the 29th acm international conference on information & knowledge management* (pp. 1115–1124).

Nguyen, M., Baker, A., Neo, C., Roush, A., Kirsch, A., & Shwartz-Ziv, R. (2024). Turning up the heat: Min-p sampling for creative and coherent llm outputs. *arXiv preprint arXiv:2407.01082*.

Perret, C. A., Johnson, A. M., McCarthy, K. S., Guerrero, T. A., Dai, J., & McNamara, D. S. (2017). Stairstepper: An adaptive remedial istart module. In *Artificial intelligence in education: 18th international conference, aied 2017, wuhan, china, june 28–july 1, 2017, proceedings 18* (pp. 557–560).

Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., ... others (2024). Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the association for computational linguistics: Naacl 2024* (pp. 1504–1518).

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, *36*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 2383–2392).

Richardson, M., Burges, C. J., & Renshaw, E. (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 193–203).

Scaria, N., Dharani Chenna, S., & Subramani, D. (2024). Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *International conference on artificial intelligence in education* (pp. 165–179).

Shuai, P., Li, L., Liu, S., & Shen, J. (2023). Qdg: A unified model for automatic question-distractor pairs generation. *Applied Intelligence*, *53*(7), 8275–8285.

Tang, L., Laban, P., & Durrett, G. (2024). Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 conference on empirical methods in natural language processing.* Association for Computational Linguistics. Retrieved from `https://arxiv.org/pdf/2404.10774`

Wang, H.-J., Hsieh, K.-Y., Yu, H.-C., Tsou, J.-C., Shih, Y. A., Huang, C.-H., & Fan, Y.-C. (2023). Distractor generation based on text2text language models with pseudo kullback-leibler divergence regulation. In *Findings of the association for computational linguistics: Acl 2023* (pp. 12477–12491).

Xu, Y., Wang, D., Yu, M., Ritchie, D., Yao, B., Wu, T., ... others (2022). Fantastic questions and where to find them: Fairytaleqa–an authentic dataset for narrative comprehension. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 447–460).

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018). Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2369–2380).