

A Neuro-symbolic reranking pipeline for level-aware learning-object recommendation: A case study in School-Principal Training

David Melo da Luz¹, Dilvan de Abreu Moreira¹, and Wesley Ferreira Maia²

¹ Universidade de São Paulo, São Paulo, Brazil

² University of California, Merced, Merced, CA, USA

Abstract. Recommending Learning Objects (LOs) in educational settings poses a challenge that conventional retrieval and RAG-inspired pipelines often overlook: beyond topical relevance, recommendations must match the learner’s proficiency level, respect prerequisites, and honor cognitive progression. We study this through continuing training for Brazilian school principals, focused on one specific competency of the Pedagogical dimension of the *Base Nacional Comum de Competências do Diretor Escolar* (BNC-Diretor; Parecer CNE/CP nº 4/2021). We investigate whether a light symbolic layer on top of a strong IR backbone can deliver pedagogical fit without sacrificing topical relevance. The pipeline combines hybrid retrieval (BM25 + multilingual dense bi-encoder fused via RRF), a multilingual cross-encoder reranker (BGE-Reranker-v2-m3), and a pedagogical reranker with six signals derived from the BNC-Diretor, a seven-level diagnostic rubric authored by a human specialist, and the platform’s controlled vocabulary. On 6,683 pedagogical-dimension LOs and 25 rubric-derived queries, the system improves over a strong hybrid + cross-encoder baseline on NDCG@10 ($\Delta = +0.041; p = 0.009$), LAS@10 ($\Delta = +0.095; p < 0.001$), and PGS@10 ($\Delta = +0.104; p < 0.001$). To mitigate circular evaluation, $\text{NDCG}_{\text{ped}}^*$ (without the LAS term) confirms the gain ($\Delta = +0.047; \text{Holm} = 0.001$). Human validation (κ up to 0.937 on 210 annotated pairs; Spearman $\rho = 0.880$ on 98 LOs) and a study on 20 real principal queries (LAS $\Delta = +0.081$; PGS $\Delta = +0.070$) indicate the gain is not an artifact of vocabulary overlap. As a side finding, on this curated-vocabulary repository BM25 outperforms the multilingual dense bi-encoder on every metric.

Keywords: RAG · Learning Object Recommendation · Pedagogical Ontology · School Management · AI for Social Good · Human Evaluation

1 Introduction

The school principal’s role is central to the quality of basic education. Research on educational policy and practice identifies pedagogical leadership as one of the school-level factors most associated with student achievement [1,2]. In Brazil, this motivated the 2021 approval, by the National Council of Education (CNE), of the *Base Nacional Comum de Competências do Diretor Escolar*

(BNC-Diretor; Parecer CNE/CP nº 4/2021), a regulatory document organized around four dimensions of school-principal practice (Political-Institutional, *Pedagogical*, Administrative-Financial, and Personal-Relational), within which ten general (transversal) competencies and seventeen specific competencies are articulated [1]. The Pedagogical dimension comprises five specific competencies (B.1–B.5 in the official text), and our case study focuses on *conducting pedagogical planning* (B.2).

Continuing training of Brazilian principals is supported by large public repositories of Learning Objects (LOs): documents, videos, tutorials, lesson plans, case studies. Yet personalised recommendation that respects each principal’s current proficiency in pedagogical planning remains an open applied problem. A beginner principal needs material different from that suited to an experienced principal who already articulates the school Political-Pedagogical Project (PPP), the National Common Curricular Base (BNCC) and innovative teaching practices. Adequate recommendation depends therefore not only on *topical relevance*, which modern IR systems handle well [3], but also on *pedagogical fit*: the material must be at the right level, cover prerequisites already mastered, and honour cognitive progression.

Retrieval-Augmented Generation (RAG) [4] is a promising architecture for this task, yet RAG pipelines are typically evaluated with topical metrics (NDCG, MRR) and do not explicitly incorporate pedagogical dimensions [5,6]. **Our claim:** a lightweight symbolic reranking layer, anchored in the BNC-Diretor and a specialist’s diagnostic rubric, improves pedagogical fit without sacrificing topical relevance. **AI for Social Good.** The Brazilian public-school system serves over 38 million students across more than 140,000 basic-education schools, and principal quality is a key lever for student learning [2]; a level-aware retrieval system aligned with the official competency framework is a step toward democratising access to specialised pedagogical knowledge. We call the approach RAG-inspired because it uses retrieval and reranking components common in RAG architectures, but it recommends LOs rather than generating free-form answers.

1.1 Contributions

1. **A three-stage neuro-symbolic retrieval and reranking pipeline** (hybrid retrieval + cross-encoder + pedagogical reranker) evaluated on a real Brazilian repository, with an effective evaluation pool of 6,683 pedagogical-dimension LOs, from a public school-principal training platform, with explicit pre-filtering by a platform-curated proxy of the Pedagogical dimension of the BNC-Diretor.
2. **A two-tier pedagogical ontology** of 25 concepts organized as a directed acyclic graph of prerequisites. Tier 1 inherits the four dimensions of the BNC-Diretor and is used as the corpus pre-filter. Tier 2 is built by a human specialist: six thematic groupings of the Pedagogical dimension (G1–G6) and 25 operational concepts whose anchors come from a seven-level diagnostic rubric and the platform’s controlled vocabulary. Tier 2 is validated

empirically (Section 6.2); Section 7.2 discusses the scope of the regulatory anchoring.

3. **An evaluation framework decoupling coupled and decoupled metrics.** We report three NDCG variants with increasing coupling to the pedagogical gold: $NDCG_{\text{top}}$ (clean), $NDCG_{\text{ped}}^*$ (composite without the optimized LAS term) and $NDCG_{\text{ped}}$ (full composite), with Holm–Bonferroni correction [7], paired bootstrap CIs over per-query differences, and two non-parametric effect sizes (Wilcoxon r and Cliff’s delta [8]).
4. **Human validation on two independent fronts:** (i) dual blind pooling of 210 (query, LO) pairs by two independent judges (Cohen’s κ up to 0.937 on topical relevance, up to 0.916 on pedagogical fit); (ii) specialist validation of the automatically inferred ontology level on a stratified sample of 98 LOs (Spearman $\rho = 0.880$, MAE = 0.54 levels, 92.9% within ± 1 level).
5. **An external-validity study with real principal queries** (N=20), in natural language and not derived from the rubric. The system keeps a significant advantage on LAS ($\Delta = +0.081$, $p < 0.001$, $r = 0.78$) and PGS ($\Delta = +0.070$, $p = 0.004$, $r = 0.89$).

We position this work as initial applied evidence: pedagogical fit improves, topical relevance does not degrade, and the effect survives multiple evaluation variants designed to detect circularity.

2 Related Work

2.1 Retrieval-Augmented Generation and Two-Stage Reranking

A bi-encoder encodes queries and documents independently for efficient semantic retrieval, whereas a cross-encoder reranker jointly scores query-document pairs to reorder a short candidate list with higher precision. Lewis et al. [4] introduced Retrieval-Augmented Generation. A modern pipeline combines (i) a retrieval stage using a dense bi-encoder [9] and/or a sparse model such as BM25 [10] with (ii) a reranking stage in which a cross-encoder reorders the top candidates with higher precision [3,11]. BGE-M3 [3] unifies dense, sparse and multi-vector retrieval in a single multilingual model. BGE-Reranker-v2-m3 [11] is widely adopted for its multilingual robustness. Reciprocal Rank Fusion [12] has become the de facto standard for hybrid pipelines [13]. Recent surveys [13,14] recommend lexical-dense hybridisation followed by cross-encoder reranking as a strong IR baseline; this is the architecture adopted as our S4 baseline.

2.2 RAG and Recommendation in Educational Settings

RAG in education has gained traction recently [5,6]. Rao and Lin [15] propose a RAG approach for MOOC recommendation; Cheng et al. [16] introduce Graph-RAG for learning-path recommendation, building prerequisite relations from textual analysis. Li et al. [17] identify prerequisites, difficulty gradients and

review strategies as core pedagogical constraints, commonly encoded via ontologies and graphs [18]. These works target individual learners; recommendation for school principals, with progression across regulatory dimensions, is under-explored in Portuguese, and to our knowledge no prior work combines a RAG pipeline with a pedagogical ontology partially anchored in an official Brazilian regulatory document.

3 Foundations: Diagnostic Rubric and Ontology

3.1 Pedagogical-Planning Rubric

The supervised input to the system is a diagnostic rubric authored by a human specialist. It classifies the principal into one of seven proficiency levels in pedagogical planning, from N1 (initial recognition of the importance of planning) to N7 (consolidated, sustainable and collaborative school culture). For each level, the rubric provides two textual feedbacks: (i) *what you already know how to do*, which reinforces the achieved level; (ii) *what you need in order to advance*, which guides toward the next level. Illustrative excerpt (N4, *to advance*): “*it will be essential to broaden the scope of diagnosis, to structure monitoring and re-planning more carefully, and to provide consistent support to teachers through clear guidelines and constructive feedback.*”

3.2 Construction of the Ontology

The operational pedagogical ontology is built in two tiers, described separately below.

Tier 1. The four dimensions of the BNC-Diretor (Político-Institucional, Pedagógica, Administrativo-Financeira, Pessoal e Relacional) are taken directly from Parecer CNE/CP nº 4/2021. The Pedagogical dimension is operationalized as a hard pre-filter: we keep LOs whose platform *Dimensions* field intersects the platform tags (School Management, Teaching and Learning Management), which the platform’s curators map to the Pedagogical dimension of the BNC-Diretor.

Tier 2. Two structures are introduced by a human specialist on top of Tier 1. (a) *Thematic groupings of the Pedagogical dimension (G1–G6)*. The specialist organizes Pedagogical-dimension content into six recurring thematic groupings, used by the reranker to compute an area-alignment signal s_{area} . The BNC-Diretor itself subdivides the Pedagogical dimension into five specific competencies (B.1–B.5); G1–G6 are independent thematic clusters observed during rubric authoring, kept for empirical utility, and Section 7.2 (7) discusses what this implies for the scope of regulatory anchoring. (b) *Twenty-five operational concepts*. Concepts are extracted by content analysis of the rubric’s “*what you need in order to advance*” feedbacks, cross-referenced with the platform’s controlled vocabulary (*Controlled terms, Keywords, Themes*). Each concept receives four attributes: (i) minimum proficiency level in the rubric (1–7); (ii) principal

grouping in G1–G6; (iii) list of prerequisites (DAG ancestors, defined by cognitive progression across levels); (iv) aliases and anchors, that is, canonical lexical forms and trigger terms for detection in the corpus.

The resulting graph is a Directed Acyclic Graph (DAG) with 25 nodes and prerequisite edges connecting adjacent levels. As an example chain, *diagnostic assessment* (N2) leads to *learning data and indicators* (N3), then to *planning monitoring* (N4) and to *re-planning* (N4); the concept *PPP* (level 6) requires *institutional goals* (N2) and *educational guidelines and policies* (N3). Empirical validation of Tier 2 is provided by dual-judge agreement on pedagogical fit (§6.2) and expert validation of inferred concept levels (§6.2).

4 Method: The Neuro-Symbolic Retrieval and Reranking Pipeline

4.1 Overview

The pipeline operates in three stages on the corpus pre-filtered by Tier 1 (§3.2). **Stage 1 (Hybrid Retrieval)**: BM25 [10] and a multilingual dense bi-encoder (paraphrase-multilingual-MiniLM-L12-v2) [9,19] are fused by Reciprocal Rank Fusion with $k = 60$ [12], returning the top 80 candidates. **Stage 2 (Cross-encoder)**: BGE-Reranker-v2-m3 [11] reorders the top 20 candidates. **Stage 3 (Pedagogical reranker)**: six signals are z-normalised on the pool from Stage 2 and combined into a final linear score. Stages 1 and 2 constitute the strong IR backbone; Stage 3 is the symbolic layer and is lightweight in practice, since every signal is computed in constant time per candidate from precomputed concept attributes, so the symbolic layer adds negligible overhead beyond the Stage 2 cross-encoder, which dominates inference cost.

4.2 The Six Signals of the Pedagogical Reranker

For each candidate d and principal at level n , we compute: s_{cross} (BGE-Reranker-v2-m3 score); s_{advance} (cosine similarity between the embedding of d and the embedding of the “to advance” text of level n); s_{preq} (proportion of prerequisites of the concepts detected in d that the principal at level n already masters, that is, DAG ancestors with level $\leq n$); s_{las} (ordinal Level Alignment Score penalising LOs far above or far below level n); s_{area} (alignment with thematic groupings G1–G6 most relevant at level n); and $s_{\text{dim}} + s_{\text{format}}$ (structured metadata: BNC-Diretor dimension and format/duration adequacy). Signal weights are given in Table 1. The final score is $s_{\text{final}} = \sum_i w_i \cdot z(s_i)$, where $z(\cdot)$ denotes z-score normalisation within the pool.

4.3 Concept Detection

Each LO is enriched with concepts detected via three cascaded methods, from most to least reliable: (i) *strong structured match*, in which the concept’s anchors

Table 1. Signals of the pedagogical reranker. Manual weights, informed by iterative ablation; a systematic grid search is left as future work (§7.2).

Signal	Definition (summary)	Weight
s_{cross}	BGE-Reranker-v2-m3 cross-encoder score	0.56
s_{advance}	Cosine similarity of d with “to advance” text at level n	0.10
s_{preq}	Fraction of prerequisites of detected concepts mastered at n	0.12
s_{las}	Ordinal Level Alignment Score between level_{LO} and n	0.10
s_{area}	Alignment with thematic groupings (G1–G6) at level n	0.06
$s_{\text{dim}} + s_{\text{format}}$	BNC-Diretor dimension and format/duration adequacy	0.06

appear in the LO’s *Controlled terms*, *Keywords*, *Themes* or *Dimensions* fields; (ii) *weak structured match*, in which aliases or variations appear in the same fields; (iii) *semantic fallback*, where the cosine similarity between the embedding of the concept name and the embedding of the LO abstract is above 0.44, limited to one concept per LO. The LO’s ontology level is computed as a confidence-weighted mean of the levels of the detected concepts. Across the 6,683 LOs in the effective pedagogical pool, 93.7% receive at least one concept; the mean is 2.66 concepts per LO (median 3.0).

4.4 Corpus

The corpus is a real repository of a Brazilian public platform for continuing school-principal training. The loaded repository contains 9,357 records after deduplication. Each LO has structured fields (type, support, duration, themes, dimensions, keywords, controlled terms) plus a narrative description. As described in §3.2 (Tier 1), initial filtering keeps only LOs whose Dimensions field intersects (School Management, Teaching and Learning Management), reducing the effective evaluation pool to 6,683 LOs.

5 Experimental Setup

5.1 Synthetic Queries

We designed 25 synthetic queries, roughly four per rubric level, derived from the *to advance* feedbacks. Each query has a target concept in the ontology. Examples: “*what is learning recomposition?*” (N1); “*how do I align planning with the school PPP?*” (N6). Queries derived from the rubric introduce construction bias, which we attack on two complementary fronts: (a) human pooling (§6.2 and §6.3); (b) an external-validity experiment with real principal queries (§6.4).

Table 2. The eight systems compared in the experiment.

System	Description
S1 BM25	BM25 only (sparse baseline)
S2 Dense	Multilingual bi-encoder only (dense baseline)
S3 Hybrid	BM25 + Dense fused by RRF (hybrid baseline)
S4 Hybrid+Cross	S3 followed by BGE-Reranker-v2-m3 (strong IR baseline)
S5 Proposed	S4 followed by the full pedagogical reranker (6 signals)
A1 No Profile	S5 without s_{advance} (profile ablation)
A2 No Ontology	S5 without s_{preq} , s_{las} , s_{area} (ontology ablation)
A3 No Metadata	S5 without s_{dim} , s_{format} (metadata ablation)

5.2 Systems Compared

Eight systems are evaluated in parallel (Table 2).

5.3 Metrics

At a high level, NDCG rewards relevant items near the top of the ranking; LAS measures alignment between the principal’s level and inferred LO level; and PGS is the fraction of top-ranked LOs pedagogically adequate for that level. We evaluate two groups of metrics on distinct gold standards [20]: (i) *topical gold*, where an LO is relevant if the query’s target concept is among the concepts detected in the LO; (ii) *composite pedagogical gold*, defined as $\text{rel}(d, q, n) = 0.45 \cdot \text{concept_hit} + 0.18 \cdot \text{prereq} + 0.17 \cdot \text{LAS} + 0.10 \cdot \text{area} + 0.10 \cdot \text{dim}$. We also report $\text{LAS}@k$ (mean Level Alignment Score in the top- k) and $\text{PGS}@k$ (fraction of the top- k with pedagogically adequate level). A methodological novelty is $\text{NDCG}_{\text{ped}}^*$: the composite gold renormalised *without* the LAS term. Since S5 optimises LAS through s_{las} , $\text{NDCG}_{\text{ped}}^*$ breaks circularity. We thus report three NDCG variants with increasing coupling: NDCG_{top} (clean) $< \text{NDCG}_{\text{ped}}^*$ (decoupled from LAS) $< \text{NDCG}_{\text{ped}}$ (full composite).

5.4 Statistical Analysis

We compute 95% bootstrap confidence intervals with 2,000 resamples over the 25 queries, both for system means and, paired, for per-query differences between S5 and each baseline. For significance we apply one-sided paired Wilcoxon tests (H_1 : $S5 > \text{baseline}$). Because S5 is compared simultaneously against four baselines (S1 to S4) on several metrics, we apply Holm–Bonferroni correction [7] to control the family-wise error rate. We report two non-parametric effect sizes: Wilcoxon $r = |Z|/\sqrt{N}$ and Cliff’s delta [8].

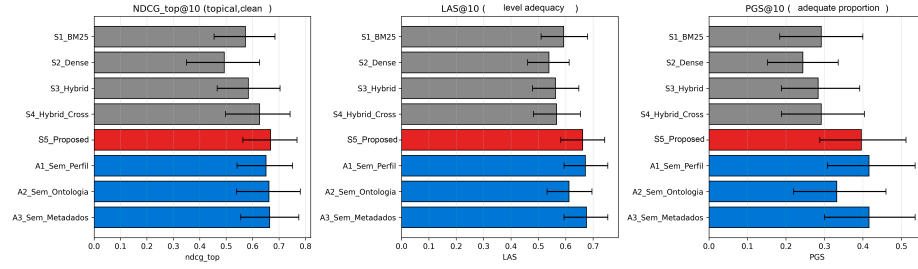
Table 3. Metrics @ $k = 10$, mean and 95% bootstrap CI (2,000 resamples, $N = 25$ queries). $NDCG^*_{\text{ped}}$ CI omitted for space.

System	$NDCG_{\text{top}}@10$	$NDCG^*_{\text{ped}}@10$	$NDCG_{\text{ped}}@10$	LAS@10	PGS@10
S1 BM25	0.573 [0.454; 0.685]	0.632	0.624 [0.552; 0.689]	0.593 [0.510; 0.680]	0.292 [0.184; 0.400]
S2 Dense	0.493 [0.350; 0.627]	0.577	0.571 [0.490; 0.650]	0.539 [0.461; 0.613]	0.244 [0.152; 0.336]
S3 Hybrid	0.584 [0.466; 0.704]	0.637	0.626 [0.554; 0.696]	0.563 [0.479; 0.648]	0.284 [0.188; 0.392]
S4 Hybrid+Cross	0.627 [0.498; 0.742]	0.662	0.647 [0.572; 0.715]	0.567 [0.482; 0.654]	0.292 [0.188; 0.404]
S5 Proposed	0.668 [0.563; 0.768]	0.709	0.704 [0.639; 0.759]	0.662 [0.582; 0.743]	0.396 [0.288; 0.512]
A1 No Profile	0.651 [0.541; 0.752]	0.699	0.696 [0.632; 0.756]	0.672 [0.594; 0.754]	0.416 [0.308; 0.536]
A2 No Ontology	0.662 [0.539; 0.781]	0.695	0.681 [0.602; 0.752]	0.612 [0.532; 0.696]	0.332 [0.220; 0.460]
A3 No Metadata	0.665 [0.555; 0.775]	0.703	0.701 [0.639; 0.761]	0.676 [0.594; 0.754]	0.416 [0.300; 0.536]

6 Results

6.1 Main Table: Three NDCGs and Adequacy Metrics

Table 3 reports metrics @10 for the eight systems, with 95% bootstrap CIs. The proposed S5 simultaneously achieves the highest of all three NDCGs, namely topical (0.668), decoupled $NDCG^*_{\text{ped}}$ (0.709) and full composite $NDCG_{\text{ped}}$ (0.704), as well as the highest LAS (0.662) and PGS (0.396).

**Fig. 1.** Main metrics per system on Block A (25 synthetic queries) with 95% bootstrap CIs.

Significance and Effect Sizes. Table 4 shows the paired S5 vs S4 contrast per metric. All contrasts except MRR are significant at the 95% level even after Holm–Bonferroni, with large effect sizes on LAS and PGS ($r > 0.8$) and a medium effect on $NDCG_{\text{top}}$ ($r \approx 0.54$). The gain on $NDCG^*_{\text{ped}}$, the decoupled variant, is $\Delta = +0.047$ (Wilcoxon $p = 0.0004$; Holm = 0.001; $r = 0.67$), which confirms that the advantage does not come from the LAS term in the gold.

6.2 Inter-Rater Agreement

Two independent judges blindly annotated 210 (query, LO) pairs, sampled by the union of the top-5 of S3, S4 and S5 plus a random sample from positions

Table 4. S5 vs S4: paired evidence with bootstrap over per-query differences, one-sided paired Wilcoxon ($H_1: S5 > S4$), Holm–Bonferroni correction across the five S5-vs-baseline contrasts, Wilcoxon r and Cliff’s δ . “n.s.” = not significant at 95%.

Metric	S5	S4	Δ	Paired 95% CI	Wilcoxon p	p_{Holm}	r / δ
$NDCG_{\text{top}}@10$	0.668	0.627	+0.041	[+0.009; +0.073]	0.0093	0.010	0.54 / 0.07
$NDCG_{\text{ped}}^*@10$	0.709	0.662	+0.047	[+0.023; +0.071]	0.0004	0.001	0.67 / 0.13
$NDCG_{\text{ped}}@10$	0.704	0.647	+0.058	[+0.034; +0.082]	0.0001	<0.001	0.73 / 0.20
LAS@10	0.662	0.567	+0.095	[+0.057; +0.142]	0.0001	<0.001	0.85 / 0.26
PGS@10	0.396	0.292	+0.104	[+0.052; +0.168]	0.0004	0.001	0.83 / 0.23
MRR@10	0.858	0.828	+0.030	[0.000; +0.087]	n.s.	n.s.	– / 0.05

Table 5. Inter-rater agreement on 210 (query, LO) pairs annotated blindly. All values lie in the “very good” range ($\kappa \geq 0.80$); quadratic-weighted κ , the variant most appropriate for ordinal scales, reaches 0.937 on relevance.

Dimension	N	κ simple	κ linear-wt.	κ quadratic-wt.
Topical relevance (0/1/2)	210	0.830	0.891	0.937
Pedagogical fit	210	0.897	0.916	–

6 to 20. Each pair was annotated on two ordinal dimensions: topical relevance on a 0/1/2 scale (irrelevant, partial, relevant) and pedagogical fit in {below, adequate, above}. Inter-rater agreement is reported in Table 5.

Expert Validation of Inferred LO Level. The specialist who authored the rubric validated the automatically inferred ontology level for a stratified sample of 98 LOs (14 per rubric level). Correlation between inferred and annotated levels is high: Spearman $\rho = 0.880$ ($p = 7.7 \times 10^{-33}$), Pearson $r = 0.891$, MAE = 0.54 levels, and 92.9% of LOs fall within ± 1 level of the annotation. Stratifying by annotator confidence band, correlation remains high even in the lowest band: $\rho = 0.897$ (low, $N=29$), $\rho = 0.814$ (medium, $N=40$), $\rho = 0.927$ (high, $N=29$). This supports the use of inferred level_{LO} in the LAS and PGS metrics.

6.3 Human Top-5 Precision per System

Table 6 reports, for each system, the fraction of LOs it contributed to the annotated pool that were classified as relevant and adequate by the human judges. We highlight *prec_adq_strict*: the fraction of LOs contributed by the system that *both* judges independently classified as “adequate” (the strictest gold label). S5 attains the highest *prec_adq_strict* (0.522), followed by S4 (0.389) and S3 (0.352), consistent with its lead on automatic LAS and PGS.

6.4 Experiment with Real Principal Queries

To test external validity, that is, whether gains persist under queries not derived from the rubric, we collected 20 real queries from three to five in-service principals, in natural language. Examples: “*what can I suggest teachers do to recover*

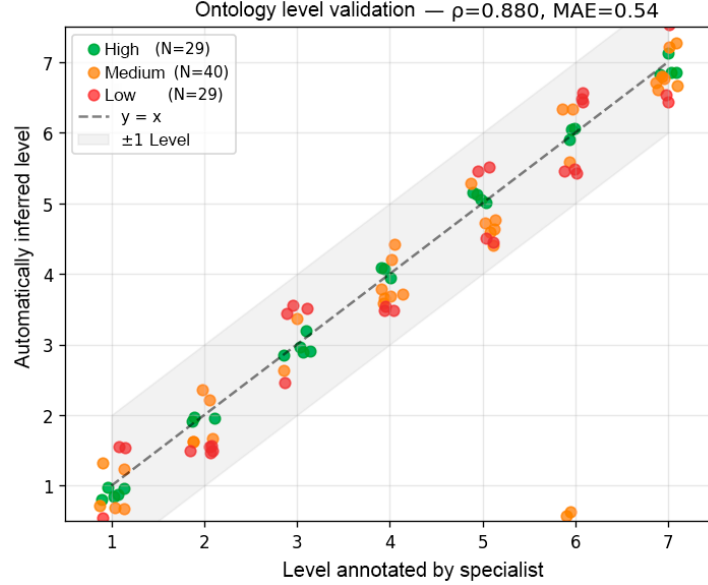


Fig. 2. Dispersion between automatically inferred LO level and expert-annotated level on 98 LOs. The grey band shows ± 1 level; the dashed line is $y = x$.

Table 6. Human top-5 precision per system over the pool of 210 dual-judged pairs.

System	N pairs	prec_rel_strict	prec_adq_strict	prec_adq_liberal	total agreement
S3	71	0.986	0.352	0.648	0.930
S4	72	0.931	0.389	0.806	0.889
S5 Proposed	67	0.955	0.522	0.746	0.910

these students without disrupting planning?” (self-declared N3); *“how do I lead the PPP revision engaging the whole school community?”* (N6). Self-reported level distribution: N1 (1), N2 (4), N3 (2), N4 (4), N5 (3), N6 (3), N7 (3). Since target concepts are only approximate in this block and proficiency levels are self-reported, we emphasise level-adequacy metrics that do not depend on the topical gold.

Table 7 compares LAS and PGS across blocks. S5 remains best (or tied), and the paired Wilcoxon S5 vs S4 on real queries confirms significance: LAS $\Delta = +0.081$, $p = 0.0004$, $r = 0.78$; PGS $\Delta = +0.070$, $p = 0.004$, $r = 0.89$. Preserving the gain in a natural-language setting not derived from the rubric is the strongest external-validity evidence in this work, although the modest sample ($N = 20$) means we treat it as preliminary.

Table 7. LAS@10 and PGS@10 on the 25 synthetic queries (rubric) vs the 20 real queries (natural-language). Ordering preserved across blocks; S5-over-S4 gains remain significant on real queries: LAS $\Delta = +0.081$ ($p < 0.001$, $r = 0.78$); PGS $\Delta = +0.070$ ($p = 0.004$, $r = 0.89$).

System	LAS synt.	LAS real	PGS synt.	PGS real
S1 BM25	0.593	0.538	0.292	0.220
S2 Dense	0.539	0.526	0.244	0.210
S3 Hybrid	0.563	0.514	0.284	0.185
S4 Hybrid+Cross	0.567	0.530	0.292	0.230
S5 Proposed	0.662	0.611	0.396	0.300
A1 No Profile	0.672	0.618	0.416	0.310
A2 No Ontology	0.612	0.580	0.332	0.275
A3 No Metadata	0.676	0.621	0.416	0.310

Table 8. Sensitivity of LAS@10 and PGS@10 to four aggregation strategies and three dispersion thresholds. S5 leads S4 in all seven variants.

Strategy	τ	LAS S4	LAS S5	PGS S4	PGS S5
weighted mean	—	0.567	0.662	0.292	0.396
median	—	0.650	0.741	0.416	0.540
max	—	0.588	0.661	0.412	0.500
min	—	0.596	0.622	0.360	0.396
weighted mean 1.5		0.639	0.730	0.380	0.494
weighted mean 1.0		0.632	0.812	0.433	0.686
weighted mean 0.5		0.798	0.923	0.678	0.878

6.5 Sensitivity, Heterogeneity, and Ablation

Sensitivity to LO-level aggregation. We tested the robustness of LAS and PGS to alternative LO-level aggregation strategies (weighted mean, median, max, min) and to a conservative mode that drops LOs with concept-level dispersion (standard deviation) above $\tau \in \{0.5, 1.0, 1.5\}$. Across all seven variants, S5 keeps LAS and PGS strictly above S4 (Table 8). The gap is largest under median aggregation (LAS $\Delta = +0.091$; PGS $\Delta = +0.124$) and under conservative filtering with $\tau = 0.5$ (LAS $\Delta = +0.125$; PGS $\Delta = +0.200$).

Heterogeneity across principal levels. Table 9 reports $NDCG_{\text{ped}}@5$ per rubric level. The proposed system beats S4 at all seven levels, with larger gains at early levels (N2: +0.17; N1: +0.10; N4: +0.09) and smaller but positive gains at intermediate and advanced levels.

Ablation of the pedagogical blocks. Table 10 shows the effect of removing each block of the pedagogical reranker. The ontology block (A2) is the only one that consistently adds value across all four metrics, with losses of 0.050 on LAS and 0.064 on PGS when removed; profile and metadata contribute marginally and

Table 9. $NDCG_{\text{ped}}@5$ per principal level. The proposed system beats S4 at all seven levels.

System	N1	N2	N3	N4	N5	N6	N7
S1 BM25	0.52	0.53	0.74	0.52	0.65	0.72	0.81
S2 Dense	0.31	0.39	0.73	0.55	0.68	0.65	0.82
S3 Hybrid	0.45	0.49	0.79	0.58	0.67	0.77	0.84
S4 Hybrid+Cross	0.46	0.49	0.81	0.60	0.71	0.75	0.84
S5 Proposed	0.56	0.66	0.83	0.69	0.79	0.77	0.90
$\Delta(\text{S5-S4})$	+0.10	+0.17	+0.02	+0.09	+0.08	+0.01	+0.06

Table 10. Ablation of the blocks of the pedagogical reranker. Values report $\Delta = \text{S5} - \text{ablation}$. For $NDCG_{\text{ped}}$ and MRR, a *positive* Δ means removing the block *improved* the metric (the block was not helpful); for LAS and PGS, a *negative* Δ means removing the block *hurt* the metric (the block was helpful). Only the ontology block (A2) contributes consistently across all four metrics.

Condition	$NDCG_{\text{ped}}$	MRR	LAS	PGS
S5 Proposed (full)	0.704	0.858	0.662	0.396
– profile structured (A1)	+0.008	+0.000	−0.010	−0.020
– ontology (A2)	+0.023	+0.010	+0.050	+0.064
– metadata (A3)	+0.004	+0.000	−0.013	−0.020

even produce small losses on LAS and PGS. We treat the ontology as the load-bearing block of the pedagogical layer; profile and metadata are kept because they are cheap, and a planned grid search (§7.2) may redistribute mass, possibly toward an even simpler symbolic layer.

6.6 Side Finding: BM25 vs Dense, and Qualitative Observation

BM25 outperforms the multilingual dense bi-encoder on every metric, with gaps between +0.048 (PGS) and +0.091 (MRR). Two properties of our setting explain this: (i) the repository’s controlled vocabulary is curated with pedagogical terminology close to the rubric, favouring lexical matching; (ii) synthetic queries are directly derived from the rubric, introducing low paraphrase [3]. We replicate this in a Brazilian educational domain and argue that, for retrieval or RAG-inspired systems over public repositories with curated metadata, BM25 should be part of the minimum baseline. Qualitatively, for “*what is learning recomposition?*” (principal at N1), S5 promotes the LO “*Pedagogical planning focused on learning recomposition*” (inferred level 2.49) into the top-5, bringing the mean top-5 level closer to the principal’s (S5: 3.31 vs S4: 3.60); the pedagogical layer does not replace the cross-encoder but refines the ranking toward level adequacy.

7 Discussion and Threats to Validity

7.1 What the Evidence Supports

The lightweight symbolic reranking layer adds measurable value to a strong IR backbone, improving pedagogical fit (LAS +0.095, PGS +0.104 on synthetic; +0.081 and +0.070 on real) without hurting topical relevance ($NDCG_{top} +0.041$). The gain survives the decoupled $NDCG_{ped}^*$, four alternative level-aggregation strategies, dual-judge pooling, and queries not derived from the rubric. Read literally, the ontology is the only block of the pedagogical reranker that consistently adds value.

7.2 Threats to Validity

We acknowledge seven limitations. **(1) Synthetic queries derived from the rubric.** The 25 main-block queries derive from *to advance* texts, introducing construction bias; we attack this with human pooling (S5 strict adequacy precision 0.522 vs 0.389 for S4) and 20 real queries from in-service principals. The $N = 20$ block is positioned as preliminary external-validity evidence, and a larger collection of in-service queries is the most informative next experiment. In the real-query block, principals’ proficiency levels were self-reported, which may introduce over- or underestimation bias. Thus, this block supports external validity for level-aware ranking, not proficiency diagnosis; future deployments should add diagnostic or expert validation. **(2 and 3) Topical gold and inferred LO level.** The topical gold rests on automatic concept detection; this is mitigated by the pooling exercise, where judges preserve the $S5 > S4 > S3$ ordering. The LO level is inferred per LO; this is mitigated by the expert-validated stratified sample of 98 LOs ($\rho = 0.880$, MAE = 0.54, 92.9% within ± 1) and the sensitivity analysis where S5 leads S4 under four aggregation strategies and three filtering thresholds. **(4) Partial circularity of the pedagogical metric.** S5 optimises LAS via s_{las} , and the composite gold includes LAS; $NDCG_{ped}^*$ (no LAS in the gold) breaks this circularity and confirms the gain ($\Delta = +0.047$; Holm = 0.001; $r = 0.67$). $NDCG_{ped}^*$ still uses automatic proxies designed jointly with the system; this point is residually vulnerable. **(5 and 6) Manual weights and single competency.** Weights were chosen manually from ablation evidence; a grid search routine on a held-out split is implemented in our codebase but was not run for this submission, and is the most immediate next experiment given the ablation pattern in Table 10. We evaluate only *conducting pedagogical planning* (B.2 of the BNC-Diretor’s Pedagogical dimension), one of seventeen specific competencies; the pipeline is competency-agnostic, so replication across B.1 and B.3 to B.5 requires only a matching rubric and concept set, with no structural change. **(7) Scope of the regulatory anchoring.** Of the two tiers of the symbolic layer, only Tier 1 (the four BNC-Diretor dimensions, used for the corpus pre-filter) is publicly verifiable against Parecer CNE/CP n^o 4/2021. Tier 2, namely the six thematic groupings G1–G6, the 25 operational concepts, and the seven-level rubric, is the specialist’s analytical contribution and has no

regulatory status; in particular, G1–G6 do not correspond to the five specific competencies B.1–B.5 that the BNC-Diretor defines under the Pedagogical dimension. The defense against the “ad hoc ontology” critique therefore rests on two kinds of evidence with distinct weights: regulatory provenance for Tier 1, and empirical validation (dual-judge agreement, expert level annotation, robustness to aggregation, real-query replication) for Tier 2.

Reproducibility. A public version of the experimental notebook is available in our GitHub repository.³ The LO repository itself is owned by a public Brazilian platform and cannot be redistributed. For this reason, the release focuses on the executable pipeline and documents the concept-detection procedure and controlled-vocabulary schema needed for independent replication.

8 Conclusion

We presented a three-stage neuro-symbolic retrieval and reranking pipeline for Learning Object recommendation in Brazilian school-principal training, with a symbolic pedagogical layer anchored in the BNC-Diretor and a specialist’s diagnostic rubric. On an effective pool of 6,683 pedagogical-dimension LOs, the system improves over a strong IR baseline on pedagogical-fit metrics (LAS +0.095, PGS +0.104) and on topical *NDCG* (+0.041), with post-Holm significance and medium-to-large effect sizes; the gain is confirmed under decoupled $NDCG^*_{\text{ped}}$, four alternative LO-level aggregations, and 20 real principal queries. Dual-judge pooling (κ up to 0.937) and specialist validation of inferred LO levels ($\rho = 0.880$) supply the empirical basis for Tier 2. Among the three blocks of the pedagogical reranker, the ontology is the load-bearing one. Future work will scale real-query collection, tune reranker weights, replicate across other BNC-Diretor competencies, and conduct a longitudinal study to measure learning gain.

References

1. Brasil, Ministério da Educação / Conselho Nacional de Educação: Base Nacional Comum de Competências do Diretor Escolar (BNC-Diretor Escolar). Parecer CNE/CP nº 4/2021, aprovado em 11 de maio de 2021. Brasília (2021). <https://www.gov.br/mec/pt-br/cne/parecer-cp-2021>
2. Leithwood, K., Harris, A., Hopkins, D.: Seven strong claims about successful school leadership. *School Leadership and Management* **28**(1), 27–42 (2008). <https://doi.org/10.1080/13632430701800060>
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., Liu, Z.: M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In: Findings of ACL 2024, pp. 2318–2335 (2024). <https://doi.org/10.18653/v1/2024.findings-acl.137>
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: NeurIPS, vol. 33, pp. 9459–9474 (2020)

³ <https://github.com/davidluz/rag-diretores-publico>

5. Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., Wang, F.L.: Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence* **8**, 100417 (2025). <https://doi.org/10.1016/j.caeai.2025.100417>
6. González-Martínez, L., et al.: Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Applied Sciences* **15**(8), 4234 (2025). <https://doi.org/10.3390/app15084234>
7. Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**(2), 65–70 (1979)
8. Meissel, K., Yao, E.: Using Cliff’s Delta as a Non-Parametric Effect Size Measure: An Accessible Web App and R Tutorial. *Practical Assessment, Research, and Evaluation* **29**(1), Article 2 (2024). <https://doi.org/10.7275/pare.1977>
9. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *EMNLP-IJCNLP 2019*, pp. 3982–3992 (2019). <https://doi.org/10.18653/v1/D19-1410>
10. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **3**(4), 333–389 (2009). <https://doi.org/10.1561/15000000019>
11. Li, C., Liu, Z., Xiao, S., Shao, Y.: Making Large Language Models A Better Foundation For Dense Retrieval. *arXiv:2312.15503v1* (2023). Model BGE-Reranker-v2-m3: BAAI, <https://huggingface.co/BAAI/bge-reranker-v2-m3>
12. Cormack, G.V., Clarke, C.L.A., Büttcher, S.: Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In: *SIGIR ’09*, pp. 758–759. ACM (2009). <https://doi.org/10.1145/1571941.1572114>
13. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H.: Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv:2312.10997* (2023)
14. Li, S., Stenzel, L., Eickhoff, C., Bahrainian, S.A.: Enhancing Retrieval-Augmented Generation: A Study of Best Practices. *arXiv:2501.07391* (2025)
15. Rao, J., Lin, J.: RAMO: Retrieval-Augmented Generation for Enhancing MOOCs Recommendations. In: *EDM Workshop Proceedings, CEUR-WS vol. 3840* (2024)
16. Cheng, X., Zhang, Z., Wang, J., Fang, L., He, C., Guan, Q., Pan, S., Luo, W.: GraphRAG-Induced Dual Knowledge Structure Graphs for Personalized Learning Path Recommendation. In: *Proceedings of the AAAI Conference on Artificial Intelligence* **40**(17), 14610–14620 (2026). <https://doi.org/10.1609/aaai.v40i17.38479>
17. Li, A., Li, Y., Gao, X.: Personalized Learning Path Recommendation Based on Knowledge Graphs: A Survey. *Electronics* **15**(1), 238 (2026). <https://doi.org/10.3390/electronics15010238>
18. Manrique, R., Pereira, B., Mariño, O.: Exploring knowledge graphs for the identification of concept prerequisites. *Smart Learning Environments* **6**, 21 (2019). <https://doi.org/10.1186/s40561-019-0104-3>
19. Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: *EMNLP 2020*, pp. 4512–4525 (2020). <https://doi.org/10.18653/v1/2020.emnlp-main.365>
20. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM TOIS* **20**(4), 422–446 (2002). <https://doi.org/10.1145/582415.582418>