CurricuLLM

Benchmarking Frontier Model Recall of Australian and New Zealand K-10 Curricula

A Comparative Evaluation of Parametric Knowledge and Retrieval-Augmented Generation

Dan Hart, Founder and CEO, CurricuLLM

February 2026

Abstract

Large language models (LLMs) are increasingly used in educational contexts, yet their factual knowledge of specific national and sub-national curricula remains largely untested.

We present a curriculum knowledge benchmark that systematically evaluates how accurately frontier LLMs can recall structured educational content from four Australian and New Zealand K-10 curriculum frameworks the Australian Curriculum v9, the Victorian Curriculum, the Western Australian Curriculum, and the New Zealand Curriculum.

Our benchmark comprises 1,700 programmatically generated questions across five categories, code-to-description mapping, description-to-code lookup, content point recall, subject-stage metadata recall, and open-ended topical questions, each designed to probe a different dimension of curriculum knowledge at the *Remember* level of Bloom's Revised Taxonomy.

We evaluate seven baseline LLMs (GPT-4.1-mini, GPT-4.1, GPT-5.2, Gemini 3 Pro, Gemini 3 Flash, Claude Sonnet 4.5, and Claude Haiku 4.5) using only their parametric knowledge, and compare them against CurricuLLM, a retrieval-augmented generation (RAG) system purpose-built for curriculum-aligned teacher support.

Responses are evaluated using an automated LLM-as-Judge pipeline (Gemini Flash 3) with category-specific rubrics, deterministic fast-path checks, and a human validation sweep confirming approximately 80% judge accuracy.

Across approximately 13,500 question-response pairs spanning four curricula, we find that no baseline LLM exceeds 41% overall accuracy, with code-specific knowledge (code-to-description, content recall) near zero for most models, while open-ended topical questions approach 80%.

CurricuLLM achieves 89% overall, outperforming the best baseline by 48 percentage points, with its largest advantages on precisely the structured curriculum queries most relevant to teacher workflows.

Cross-curriculum analysis reveals that all models perform best on the New Zealand Curriculum (which lacks outcome codes) and worst on the Victorian Curriculum, suggesting performance correlates with question specificity and training data prevalence.

We situate this work within Bloom's Taxonomy and argue that curriculum recall represents a necessary but insufficient condition for effective AI-assisted teaching. We outline a research agenda for future benchmarks targeting higher-order cognitive skills, comprehension, application, analysis, and evaluation, to more fully assess the pedagogical utility of LLMs.

Keywords: large language models, curriculum alignment, educational benchmarking, retrieval-augmented generation, Bloom's Taxonomy, K-10 education, Australian Curriculum, New Zealand Curriculum

1. Introduction

The rapid integration of large language models into educational workflows has created an urgent need to understand what these models know about the curricula that structure K-10 teaching and learning. Teachers across Australia and New Zealand increasingly turn to AI assistants for lesson planning, resource creation, and curriculum mapping [19, 24, 25], yet the factual accuracy of these tools with respect to specific national and state curriculum frameworks has not been systematically evaluated.

Existing LLM benchmarks in education primarily test general academic knowledge through standardised tests drawn from United States contexts. MMLU [1] spans 57 academic subjects but at a tertiary level with no curriculum-specific metadata. The ARC dataset [2] uses US grade-school science questions. AGIEval [4] draws on standardised exams from multiple countries but focuses on examination performance rather than curriculum structure. None of these benchmarks test whether an LLM can identify the codes, subjects, year levels, content points, or strands associated with learning outcomes in a particular jurisdiction's curriculum, the kind of structured knowledge that curriculum-aligned AI tools must possess.

This gap matters for several reasons. First, Australia and New Zealand employ distinct, overlapping curriculum frameworks. The Australian Curriculum v9 provides a national framework, but states such as Victoria and Western Australia maintain their own curriculum documents with jurisdiction-specific codes, strands, and content elaborations. New Zealand's curriculum differs further in structure, terminology, and pedagogical philosophy. An AI tool that conflates these frameworks, or that hallucinates plausible-sounding but incorrect curriculum content, could mislead teachers and undermine trust in AI-assisted planning.

Second, the distinction between parametric knowledge (what an LLM has memorised from its training data) and retrieval-augmented knowledge (what an LLM can access through external retrieval at inference time) is critical for curriculum applications. Australian and New Zealand curricula likely constitute long-tail knowledge relative to the predominantly English-language web corpora on which frontier LLMs are trained [32]. Retrieval-augmented generation (RAG) systems [11] offer a promising approach to grounding LLM responses in authoritative curriculum data, but the magnitude of improvement over parametric knowledge alone has not been quantified for this domain.

This challenge is compounded by the fact that all curricula under test are currently undergoing multi-year revision processes. The Australian Curriculum transitioned from Version 8 to Version 9 beginning in 2022, with state and territory adoption timelines varying. New Zealand is refreshing its curriculum through the Te Mātaiaho framework. As a result, LLMs whose training data includes earlier curriculum versions may produce answers that were once correct but are now outdated, and parametric knowledge of superseded content may actually reduce accuracy rather than improve it on questions about current curricula.

Third, while curriculum recall, the ability to accurately reproduce factual information about curriculum content, is a necessary foundation, it represents only the lowest level of Bloom's Revised Taxonomy [16, 17]. Effective AI teaching assistants must also demonstrate comprehension (explaining why a content descriptor is taught at a particular year level), application (generating assessment tasks aligned to specific outcomes), analysis (identifying cross-curricular connections), and evaluation (critiquing a lesson plan for curriculum

alignment). Understanding baseline recall performance is a prerequisite for designing benchmarks that test these higher-order capabilities.

In this paper, we make the following contributions:

1. We present a curriculum knowledge benchmark comprising approximately 1,700 questions (after data curation) across five categories, generated from structured curriculum databases covering four Australian and New Zealand K-10 frameworks.
2. We evaluate seven frontier LLMs and one RAG system (CurricuLLM), providing the first systematic comparison of parametric versus retrieval-augmented curriculum knowledge.
3. We describe a robust automated evaluation pipeline combining an LLM-as-Judge approach with deterministic fast-path optimisations, category-specific rubrics, and human validation.
4. We situate our benchmark within Bloom's Revised Taxonomy and propose a research agenda for evaluating higher-order curriculum reasoning capabilities.

2. Related Work

2.1 LLM Evaluation in Education

The evaluation of LLMs on educational tasks has evolved rapidly. The Massive Multitask Language Understanding (MMLU) benchmark [1] established the paradigm of testing LLMs across dozens of academic subjects, revealing that model scale correlates with multitask accuracy. The AI2 Reasoning Challenge (ARC) [2] specifically targeted grade-school science, while SciQ [3] demonstrated crowdsourced question generation for science education. AGIEval [4] expanded evaluation to real standardised exams including college entrance tests, finding GPT-4 achieved 95% on SAT Math. The GPT-4 Technical Report [5] further demonstrated near-human performance on professional and academic examinations across diverse domains.

More recently, Rodrigues et al. [6] evaluated GPT-4 on 7,380 open-ended high-school questions categorised by Bloom's Taxonomy level and Item Response Theory difficulty, finding performance comparable to native-speaking students. The OpenLearnLM benchmark [30] introduced a unified framework of over 124,000 items for evaluating educational LLMs across knowledge, skill, and attitude dimensions. Lelièvre et al. [31] benchmarked pedagogical knowledge specifically, testing 97 models on 920 multiple-choice questions drawn from teacher training examinations. Henkel et al. [29] examined LLMs' ability to grade K-12 student responses in Science and History, finding GPT-4 achieved near-human inter-rater reliability.

However, all of these benchmarks test general academic knowledge or pedagogical reasoning. None evaluate whether LLMs can accurately recall the specific structure, codes, content descriptors, and organisational metadata of national or sub-national curriculum frameworks, the kind of domain-specific knowledge required for curriculum-aligned AI tools.

2.2 LLM-as-Judge Methodology

The use of LLMs as automated evaluators has been validated extensively. Zheng et al. [7] introduced the LLM-as-Judge framework with MT-Bench and Chatbot Arena, demonstrating

that GPT-4 achieves over 80% agreement with human evaluators, matching inter-annotator agreement rates. Liu et al. [8] proposed G-Eval, using chain-of-thought prompting for evaluation, achieving state-of-the-art correlation with human judgments. Wang et al. [9] identified systematic position bias in LLM evaluation and proposed calibration strategies. Li et al. [10] provide a survey of the LLM-as-Judge paradigm, covering scoring methods, biases, and mitigation strategies.

Our benchmark adopts the LLM-as-Judge approach using Gemini Flash 3 as the evaluator, with several mitigations for known biases: we use category-specific rubrics to reduce subjectivity, implement deterministic fast-path checks that resolve unambiguous cases without invoking the judge, and conduct a human validation sweep that confirmed approximately 80% judge accuracy. We note that Gemini Flash 3 also serves as one of the models under test; this dual role and its implications are discussed in Section 5.6.

2.3 Retrieval-Augmented Generation in Education

Retrieval-augmented generation (RAG), introduced by Lewis et al. [11], combines parametric language model knowledge with non-parametric retrieval from external knowledge bases. The paradigm has been widely adopted in education. Li et al. [12] provide a survey of RAG in educational applications, categorising uses across interactive learning systems, content generation, and institutional deployment. Dong [13] demonstrated that knowledge graph-enhanced RAG (KG-RAG) improved assessment scores by 35% in an AI tutoring context, highlighting the value of structured knowledge representation. Han et al. [14] showed that RAG-based approaches outperformed zero-shot and chain-of-thought strategies for automated assessment of tutoring practices.

CurricuLLM, the RAG system evaluated in this benchmark, is a production AI assistant purpose-built for Australian and New Zealand teachers. It leverages retrieval augmentation over structured curriculum data to ground its responses in authoritative curriculum content, providing an end-to-end comparison point against baseline LLMs operating solely from parametric knowledge.

2.4 Parametric vs. Retrieval-Augmented Knowledge

The distinction between what LLMs know from training (parametric knowledge) and what they can access through retrieval (non-parametric knowledge) is central to our benchmark design. Mallen et al. [32] demonstrated that LLMs struggle with less popular factual knowledge while retrieval augmentation provides the greatest benefit for long-tail topics. This finding is directly relevant: Australian and New Zealand curricula are likely underrepresented in LLM training corpora relative to US and UK educational content.

Xie et al. [33] conducted a study of LLM behaviour under knowledge conflicts between parametric memory and external evidence, finding that models exhibit confirmation bias when evidence partially aligns with existing knowledge. Xu et al. [34] surveyed three types of knowledge conflicts, context-memory, inter-context, and intra-memory, providing a framework for understanding how retrieval-augmented systems handle curriculum content that may differ from what the model learned during training. Longpre et al. [35] established foundational methods for studying entity-based knowledge conflicts in question answering.

2.5 Bloom's Taxonomy and LLM Evaluation

Bloom's Taxonomy [15], revised by Anderson and Krathwohl [16, 17], provides a six-level cognitive hierarchy: Remember, Understand, Apply, Analyse, Evaluate, and Create. This framework has been increasingly applied to LLM evaluation. Huber and Niklaus [18] mapped existing LLM benchmarks to Bloom's levels and found that current evaluation is heavily biased toward lower-order cognitive skills, with higher levels (Evaluate, Create) significantly underrepresented. This finding directly motivates both our current benchmark, which explicitly targets the Remember level, and our proposed future work on higher-order curriculum reasoning.

Kasneci et al. [19] discussed the opportunities and challenges of LLMs in education from both student and teacher perspectives, including the limitation that current LLMs may lack higher-order reasoning capabilities necessary for effective pedagogical support.

2.6 AI in Australian and New Zealand Education

The Australian Curriculum v9 [20], endorsed in 2022, provides the national K-10 framework against which state curricula are aligned. The Australian Government's Framework for Generative AI in Schools [21], released in 2023, established six guiding principles for responsible AI use in K-12 education, signalling institutional commitment to AI integration. The New Zealand Curriculum [22] and accompanying Ministry of Education guidance on generative AI [23] provide the policy context for NZ-specific evaluation.

Empirical evidence of AI adoption in these jurisdictions is emerging. Coblenz et al. [24] found that 69% of New Zealand primary school teachers use AI weekly for lesson planning and assessment. Bower et al. [25] examined priorities identified by senior Australian education policy makers regarding generative AI, finding that risk management, teacher education, and system leadership were paramount concerns. These findings underscore the urgency of evaluating AI tools' curriculum knowledge, teachers are already relying on LLMs for curriculum-related tasks, yet no benchmark exists to assess whether these tools provide accurate curriculum information.

2.7 Curriculum-Aligned AI Tools

Several recent works address the challenge of aligning LLM outputs with educational standards. Imperial et al. [26] introduced a retrieval-based framework that improved standard alignment accuracy by 45-100% when guiding LLMs to generate content aligned with Common European Framework of Reference for Languages and Common Core standards. Liu et al. [27] incorporated curriculum components grounded in the Next Generation Science Standards to generate grade-appropriate educational content. These works demonstrate growing interest in curriculum-AI alignment but focus primarily on content generation rather than knowledge evaluation.

2.8 Benchmark Design

Best practices for LLM benchmark design have been formalised in several works. Liang et al. [36] established the HELM methodology for holistic evaluation, introducing taxonomic approaches covering accuracy, calibration, robustness, fairness, and efficiency. The problem of benchmark data contamination, where LLM training data includes benchmark questions, has been extensively studied by Xu et al. [37] and Deng et al. [38], the latter finding that

GPT-4 could guess missing MMLU options at 57% exact match rate. White et al. [39] proposed LiveBench, using frequently updated questions to resist contamination. Chang et al. [40] provide a survey of LLM evaluation methodology, covering task design, metrics, and evaluation protocols. These works inform our benchmark design decisions, particularly regarding question generation, evaluation methodology, and threats to validity.

3. Methodology

This section describes our benchmark design, including the curricula under test, models evaluated, question generation pipeline, and evaluation methodology.

3.1 Overview

Our benchmark follows a three-phase pipeline: (1) structured questions are programmatically generated from curriculum content databases, supplemented by LLM-generated open-ended questions; (2) each question is sent to every model under test via API; and (3) responses are evaluated against ground truth answers, returning a binary PASS/FAIL verdict. Where possible, evaluation uses deterministic matching, including exact code matching, year-band matching, and language subject matching, bypassing the LLM judge entirely. For cases that cannot be resolved deterministically, an independent LLM judge (Gemini Flash 3) evaluates the response with category-specific rubrics. A human validation sweep of a random sample confirmed approximately 80% judge accuracy.

3.2 Curricula Under Test

The benchmark covers four curriculum frameworks spanning national and state jurisdictions across Australia and New Zealand:

Curriculum	Key	Jurisdiction
Australian Curriculum v9	aus-v9	Australia (National)
Victorian Curriculum	vic	Victoria, Australia
Western Australian Curriculum	wa	Western Australia
New Zealand Curriculum	nz	New Zealand

It is important to note that all four curricula are currently undergoing multi-year revision processes. The Australian Curriculum transitioned from Version 8 to Version 9 (endorsed 2022), with state adoption occurring on varying timelines. New Zealand is developing its Te Mātaiaho curriculum refresh. This means that LLMs trained on web data from different time periods may have internalised different, and potentially conflicting, versions of curriculum content. Our benchmark uses the current (2026) state of each curriculum as ground truth, meaning models with outdated training data may be penalised for answers that were correct under previous versions.

3.3 Models Under Test

We evaluate models in two categories. Baseline LLMs are tested via their native APIs with only parametric knowledge, no retrieval augmentation or curriculum-specific context.

CurricuLLM is tested as a complete system through its production API, reflecting its full capabilities including retrieval augmentation over curriculum data.

Model ID	Model Name	Provider	Category
gpt-4.1-mini	GPT-4.1-mini	OpenAI	Baseline + Judge
gpt-4.1	GPT-4.1	OpenAI	Baseline
gpt-5.2	GPT-5.2	OpenAI	Baseline
gemini-3-pro	Gemini 3 Pro	Google (Vertex AI)	Baseline
gemini-3-flash	Gemini 3 Flash	Google (Vertex AI)	Baseline
sonnet-4.5	Claude Sonnet 4.5	Anthropic	Baseline
haiku-4.5	Claude Haiku 4.5	Anthropic	Baseline
CurricuLLM	CurricuLLM	CurricuLLM	RAG-augmented

Baseline models receive a minimal system prompt: “You are a helpful assistant for teachers using the [Curriculum Name]. Answer questions about curriculum content accurately and concisely.” This tells the model which curriculum is being tested without providing any curriculum content, ensuring the benchmark measures parametric knowledge only. All models (baseline and judge) are run at temperature 1.0.

CurricuLLM is tested end-to-end through its production API: for each question, a new conversation is created and the question is sent as a user message, with the full response collected. This ensures results reflect the system’s real-world capabilities as experienced by teachers, rather than any isolated component.

3.4 Question Generation

Questions are generated across five categories, each testing a different dimension of curriculum knowledge. Four categories are generated programmatically from structured curriculum data; one is generated by an LLM with curriculum context.

3.4.1 Question Categories

Code-to-Description (code_to_description). Given an outcome code (e.g., AC9M3A01), the model must produce the corresponding outcome description. This tests precise recall of curriculum code-description mappings.

Description-to-Code (description_to_code). Given an outcome description, the model must produce the corresponding code. This tests reverse lookup capability and is evaluated by checking for the presence of an acceptable code in the response.

Content Recall (content_recall). Given an outcome code, the model must name a content point or elaboration associated with that outcome. This tests depth of knowledge beyond the top-level description.

Subject-Stage Recall (subject_stage_recall). This category tests curriculum organisational knowledge through three sub-types: identifying the subject or year level for a given code; naming an outcome in a specified subject at a specified year level; or describing content taught in a subject-year combination.

Vague/Topical (vague_topical). Open-ended, teacher-style questions generated by Gemini 3 Pro, such as “Name one outcome that teaches fractions” or “Which subject covers persuasive writing?”

3.4.2 Question Distribution and Sampling

The target is approximately 500 questions per curriculum, though the actual count varies depending on the categories available for each framework. The distribution across categories is adaptive: curricula with both outcome codes and content points distribute questions evenly across all five categories; curricula lacking codes or content points exclude those categories entirely rather than redistributing quotas. For example, the New Zealand Curriculum, which does not have outcome codes, generates only subject-stage recall and vague/topical questions, resulting in a smaller overall question set. Following automated generation, a human review removes questions deemed too vague or unanswerable, further reducing the final count.

3.4.3 Answer Expansion and Post-Processing

To reduce false negatives during evaluation, acceptable answer lists are expanded through several mechanisms. Cross-subject similarity matching using Jaccard similarity (≥ 0.7 on word sets) identifies templated outcomes that share descriptions across language subjects (e.g., French, Japanese, Italian) but have different codes. Same-subject, all-stages expansion captures topics that recur across year levels. Content recall answers are augmented with 3-5 LLM-generated paraphrases (Gemini Flash 3). Vague/topical answers are expanded through a curriculum-wide scan using code-based expansion, topic keyword search, and stage/year sweeps.

3.5 Model Execution

Baseline models receive questions as independent requests (no conversational history) via their native APIs. CurricuLLM is tested through its production API with each question sent as a new conversation, ensuring realistic end-to-end conditions.

3.6 Evaluation

Each response is evaluated by Gemini Flash 3 acting as an independent judge. The judge receives the question text, category, expected answer, additional acceptable answers, the model’s response, and a curriculum database lookup section when applicable. We note that Gemini Flash 3 is both judge and one of the models under test; implications of this dual role are discussed in Section 5.6. A human validation sweep of a random sample of judge verdicts confirmed approximately 80% accuracy, consistent with expected LLM-as-judge reliability for factual knowledge evaluation tasks.

3.6.1 Category-Specific Evaluation Criteria

The judge applies different standards by category. Code-to-description questions require semantic equivalence; paraphrasing is acceptable if core meaning matches. Description-to-code questions accept any code whose looked-up description is semantically equivalent to the question. Content recall accepts any valid content point with paraphrasing. Subject-stage recall requires exact match for metadata queries but accepts any valid outcome for naming tasks, with year-band and language subject matching accommodations. Vague/topical questions apply the most lenient criteria: any response demonstrating

genuine, accurate curriculum knowledge passes, even if not in the provided acceptable answer list.

3.6.2 Fast-Path Optimisations

Five deterministic fast-paths bypass the LLM judge for efficiency: empty responses are automatically removed from the test; acceptable code matches in the response trigger automatic pass; year-band matching resolves stage questions; stage matching handles topical temporal questions; and language subject matching accommodates the Languages learning area hierarchy. A code lookup augmentation step extracts outcome-code-like strings from responses and verifies them against the curriculum database, providing the judge with concrete evidence for alternative codes.

3.7 Human Validation

Following automated evaluation, a human validation sweep was conducted to calibrate confidence in the LLM judge's verdicts. A random sample of judge verdicts was independently reviewed by a human assessor, who re-evaluated sampled responses against the ground truth answers and compared their verdicts with the automated judge's decisions. Results aligned with approximately 80% accuracy, consistent with expected LLM-as-judge reliability for factual knowledge evaluation tasks. This human validation serves as a calibration check confirming the automated evaluation produces results within acceptable bounds, rather than a full re-evaluation of all responses.

4. Results

We present results from the complete benchmark across all four curricula: Australian Curriculum v9 (500 questions), Victorian Curriculum (500), Western Australian Curriculum (500, no content recall category), and New Zealand Curriculum (300, subject-stage recall and vague/topical only). After data curation, excluding connection errors, the final dataset comprises approximately 1,690 evaluated question-response pairs per baseline model and 1,589 for CurricuLLM. Throughout this section, we refer to the seven models tested without retrieval augmentation as "baseline LLMs" and to CurricuLLM as the "RAG system."

4.1 Overall Pass Rates

Table 1 reports overall pass rates for all eight models across all curricula combined.

Model	Category	Passed	n	Pass Rate
CurricuLLM	RAG system	1,409	1,589	**88.7%**
Gemini 3 Pro	Baseline	693	1,690	41.0%
Gemini 3 Flash	Baseline	645	1,691	38.1%
GPT-5.2	Baseline	613	1,691	36.3%
GPT-4.1	Baseline	582	1,691	34.4%
Sonnet 4.5	Baseline	554	1,688	32.8%
GPT-4.1-mini	Baseline	454	1,691	26.8%
Haiku 4.5	Baseline	418	1,690	24.7%

CurricuLLM achieved an overall pass rate of 88.7%, outperforming the best baseline LLM (Gemini 3 Pro, 41.0%) by 47.7 percentage points. Among baseline models, Gemini 3 Pro led (41.0%), followed closely by Gemini 3 Flash (38.1%) and GPT-5.2 (36.3%). A middle tier comprised GPT-4.1 (34.4%) and Sonnet 4.5 (32.8%), with the smaller models GPT-4.1-mini (26.8%) and Haiku 4.5 (24.7%) trailing. No baseline model exceeded 41% overall pass rate, indicating that curriculum-specific factual knowledge is sparse in the parametric memory of all frontier LLMs tested.

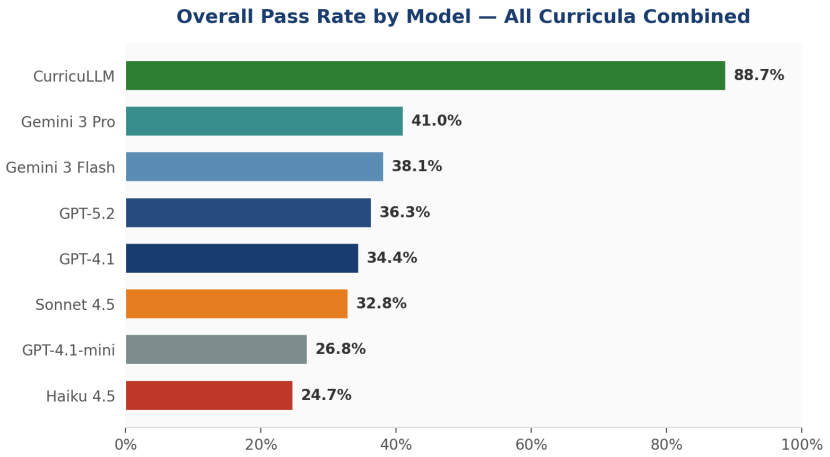

Figure 1. Overall pass rates by model across all four curricula combined.

4.2 Performance by Question Category

Table 2 disaggregates pass rates by question category across all curricula, revealing stark differences in performance across the five dimensions of curriculum knowledge tested.

Model	Code→Desc	Content	Desc→Code	Subj/Stage	Vague
CurricuLLM	82.5%	73.2%	88.8%	90.9%	98.4%
Gemini 3 Pro	6.0%	16.5%	33.5%	45.3%	80.4%
Gemini 3 Flash	4.7%	12.5%	35.7%	38.2%	77.3%
GPT-5.2	8.5%	3.5%	27.7%	42.0%	73.5%
GPT-4.1	4.4%	6.0%	29.8%	35.4%	73.0%
Sonnet 4.5	0.0%	4.0%	28.3%	34.9%	72.3%
GPT-4.1-mini	0.9%	1.5%	23.7%	19.3%	68.3%
Haiku 4.5	0.0%	1.0%	24.4%	13.2%	66.4%

The category-level results reveal a dramatic gradient of difficulty for baseline LLMs. Code-to-description and content recall proved essentially impossible for most baselines: two models scored 0.0% on code-to-description, and the highest baseline achieved just 8.5% (GPT-5.2). Content recall showed a similar pattern, with most baselines below 7%. These results indicate that baseline LLMs have not memorised the mapping between curriculum outcome codes and their descriptions to any meaningful degree.

Description-to-code performance was notably higher (23.7-35.7% across baselines) than code-to-description (0.0-8.5%). Subject-stage recall showed moderate performance

(13.2-45.3% across baselines), suggesting LLMs have partial knowledge of curriculum organisation.

Vague/topical questions showed the highest baseline performance (66.4-80.4%), with Gemini 3 Pro achieving 80.4%. This demonstrates that LLMs retain reasonable conceptual knowledge of what topics are taught at various year levels, even when they cannot recall specific codes or content descriptors. CurricuLLM achieved 98.4% on vague/topical questions, substantially exceeding all baselines. Across all curricula, the gradient from 3.5% (code-to-description baseline average) to 73.0% (vague/topical baseline average) confirms that curriculum recall is a spectrum: models know the broad shape of curriculum content but lack the precise details.

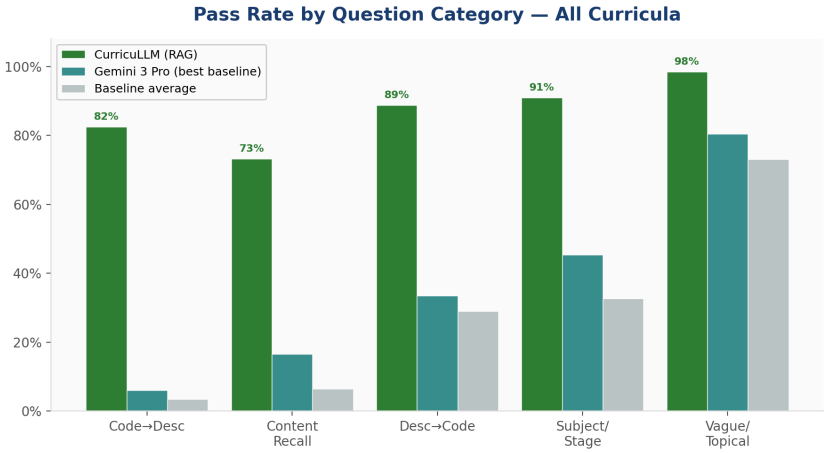

Figure 2. Pass rates by question category: CurricuLLM (RAG) vs. best baseline (Gemini 3 Pro) vs. baseline average, all curricula combined.

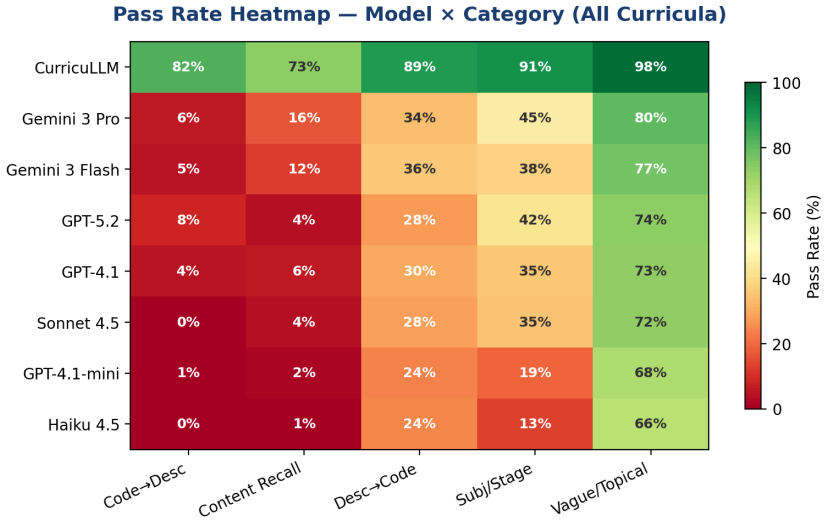

Figure 3. Pass rate heatmap across all models and question categories. The red-to-green gradient reveals the sharp contrast between code-dependent categories (near-zero for baselines) and open-ended categories.

4.3 CurricuLLM vs. Baseline LLMs

CurricuLLM's retrieval augmentation produced its largest advantages on the code-dependent categories: +76.5 percentage points over the best baseline on

code-to-description (82.5% vs. 6.0%), +56.7 on content recall (73.2% vs. 16.5%), and +53.1 on description-to-code (88.8% vs. 35.7%). The advantage narrowed on subject-stage recall (+45.6 pp) but remained substantial even on vague/topical questions (+18.0 pp: 98.4% vs. 80.4%).

The 98.4% pass rate on vague/topical questions is particularly notable: CurricuLLM's retrieval pipeline appears to provide highly effective grounding for open-ended teacher queries, anchoring responses to specific curriculum content while still covering the breadth of valid answers.

4.4 Model Size and Performance

We classify baseline models into two tiers based on each provider's product positioning: small models (Gemini 3 Flash, GPT-4.1-mini, Haiku 4.5) represent each provider's lighter, cost-optimised offering, while large models (Gemini 3 Pro, GPT-4.1, GPT-5.2, Sonnet 4.5) represent the flagship or full-capability tier. Exact parameter counts are not published, so this classification reflects market positioning rather than architectural detail.

Within each family, the expected size gradient held: Gemini 3 Pro (41.0%) outperformed Gemini 3 Flash (38.1%), GPT-5.2 (36.3%) and GPT-4.1 (34.4%) both outperformed GPT-4.1-mini (26.8%), and Sonnet 4.5 (32.8%) outperformed Haiku 4.5 (24.7%). However, the cross-family variation was substantially larger than the within-family gradient. Most notably, the small-tier Gemini 3 Flash (38.1%) outperformed three of the four large-tier models, GPT-4.1 (34.4%), GPT-5.2 (36.3%), and Sonnet 4.5 (32.8%). This suggests that model family, likely reflecting differences in training data composition and curation, is a stronger predictor of curriculum recall than model size class.

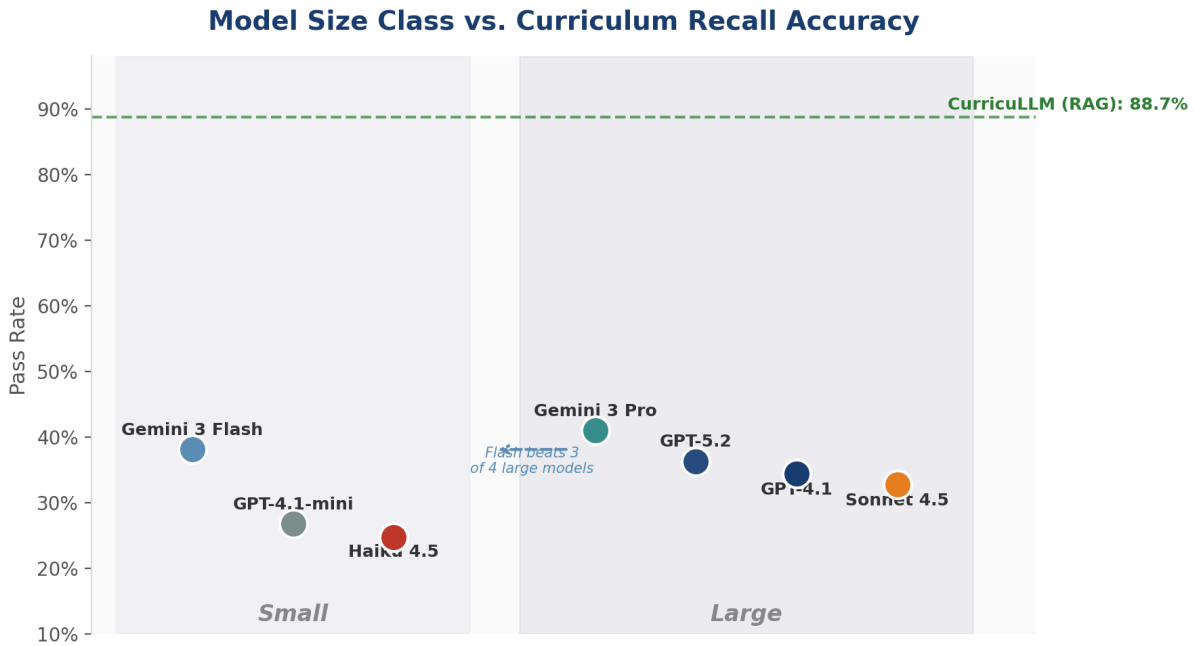

Figure 4. Model tier versus curriculum recall accuracy. Gemini models outperform larger models from other families, suggesting training data composition matters more than scale for domain-specific recall. CurricuLLM reference line (dashed) shows the ceiling enabled by retrieval augmentation.

4.5 Cross-Curriculum Analysis

Performance varied substantially across curricula. All models performed best on the New Zealand Curriculum (baseline range: 43.2-66.3%, CurricuLLM: 98.4%) and worst on the Victorian Curriculum (baseline range: 14.8-31.8%, CurricuLLM: 88.3%). The Australian Curriculum v9 (17.4-36.8%) and Western Australian Curriculum (33.1-44.9%) fell in between.

The New Zealand Curriculum's higher pass rates are explained by its question composition: NZ lacks outcome codes, so only subject-stage recall and vague/topical questions are tested, the two easiest categories for all models. This structural difference means NZ results are not directly comparable to the three Australian curricula. Among the Australian curricula, the Victorian Curriculum was consistently the hardest for baselines, possibly reflecting less web presence for Victorian-specific curriculum documents compared to the national Australian Curriculum.

The Western Australian Curriculum showed higher baseline pass rates than the Australian Curriculum v9 or Victorian Curriculum across most models. This is partly structural, WA lacks content recall questions (one of the hardest categories for baselines), so its overall rate is computed across four categories rather than five. However, WA also showed genuinely higher baseline performance on description-to-code questions, every baseline model scored above 60% on WA description-to-code, compared to under 22% for the same category on the Australian Curriculum v9. This suggests that WA's SCSA-style outcome codes are better represented in LLM training data than v9 codes, possibly because WA curriculum documents have been publicly accessible online for longer.

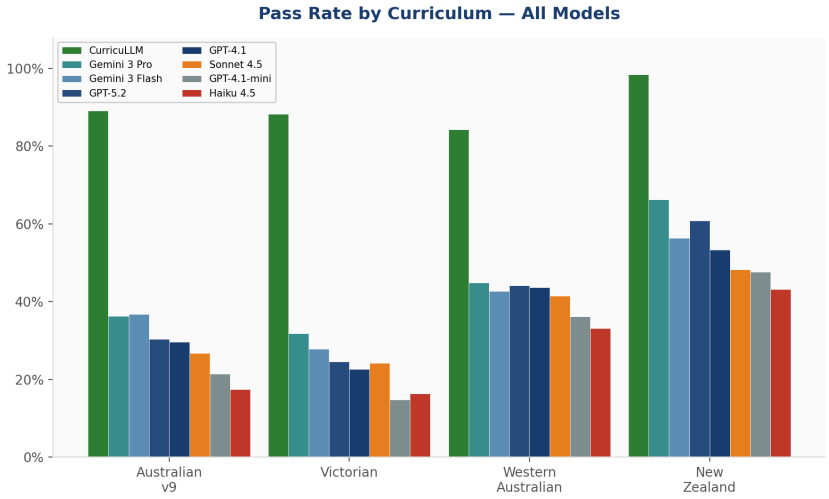

Figure 5. Pass rates by curriculum for all models. New Zealand's higher scores reflect its simpler question composition (no code-based categories). The Victorian Curriculum is consistently the hardest for baseline models.

5. Discussion

5.1 Parametric Knowledge of Regional Curricula

The results strongly support the hypothesis that Australian and New Zealand curriculum content constitutes long-tail knowledge for frontier LLMs [32]. The best baseline model achieved only 41.0% overall pass rate, and on the most demanding categories (code-to-description, content recall) baselines peaked at just 8.5% and 16.5% respectively. This stands in sharp contrast to the performance of the same model families on US-centric educational benchmarks: GPT-4 achieves 95% on SAT Math [4, 5], yet GPT-5.2 achieves just 36.3% on curriculum recall. The knowledge deficit is not one of general educational competence but of domain-specific factual content.

The cross-family comparison is illuminating. Gemini 3 Pro led all baselines (41.0%), followed by Gemini 3 Flash (38.1%). The Gemini advantage was concentrated in the subject-stage recall and vague/topical categories, suggesting Google's training data pipeline may include greater representation of Australian educational content, perhaps through indexing of Australian government and education authority websites, though this remains speculative.

5.2 Refusal vs. Confabulation in Baseline Models

A distinctive failure pattern emerged from the two Anthropic models. Haiku 4.5 contained explicit refusal language, typically "I don't have access to a specific database of Australian Curriculum outcome codes", in 49.2% of all questions, while Sonnet 4.5 did so in 23.1%. By contrast, OpenAI and Google models produced zero, instead always attempting an answer regardless of confidence. The refusal rate was most pronounced on code-to-description questions, where Haiku 4.5 refused 96.2% of the time and Sonnet 4.5 refused 95.3%, both models correctly recognising that they lacked the specific knowledge being tested. This creates an important interpretive distinction: Anthropic models' lower pass rates partly reflect a choice to decline rather than guess, while GPT and Gemini models' failures are almost entirely confident but incorrect responses. From a teacher's perspective, a model that says "I don't know, check the ACARA website" is arguably less harmful than one that confidently provides the wrong curriculum code. This pattern may reflect Anthropic's approach to model calibration, where models are trained to express uncertainty rather than confabulate, though it significantly penalises them under a binary PASS/FAIL evaluation scheme.

5.3 The Value of Retrieval Augmentation

CurricuLLM's 47.7-percentage-point advantage over the best baseline decisively demonstrates the value of retrieval augmentation for curriculum-specific applications. The advantage was most dramatic on precisely the categories that matter most for teacher-facing tools: code-to-description (+76.5 pp) and description-to-code (+53.1 pp) are the operations teachers perform when mapping resources to curriculum outcomes or building scope-and-sequence documents.

CurricuLLM's 98.4% pass rate on vague/topical questions, exceeding all baselines by at least 18 percentage points, demonstrates that retrieval augmentation can enhance even open-ended queries by grounding responses in authoritative curriculum content. This aligns with the knowledge conflict literature [33, 34]: when retrieval provides high-quality, relevant evidence, it resolves rather than exacerbates uncertainty. CurricuLLM's 88.7% overall

accuracy across nearly 1,600 questions indicates that retrieval augmentation provides robust, scalable access to curriculum knowledge that is fundamentally absent from parametric memory.

5.4 Baseline Configuration and Web-Augmented Models

Our benchmark evaluates baseline LLMs using only their parametric knowledge, that is, the knowledge encoded in model weights during training, without access to external tools or information sources. In practice, many consumer-facing deployments of these models now include web search integration, allowing the model to retrieve and cite external sources in real time. We did not test web-augmented configurations, and it is plausible that models with search enabled would perform better on curriculum recall tasks, particularly for questions about publicly accessible curriculum documents.

However, web search augmentation introduces its own risks during periods of curriculum transition. All four curricula tested in this benchmark are currently undergoing multi-year revision processes, meaning that web sources may contain a mixture of current and superseded content, often without clear version labelling. A web-augmented model that retrieves and confidently cites an outdated curriculum document may be more harmful than one that simply declines to answer, as the response carries the authority of a cited source while delivering incorrect information. This is precisely the failure mode that purpose-built retrieval systems like CurricuLLM are designed to avoid: by grounding retrieval in a curated, version-controlled curriculum database rather than the open web, the risk of temporal contamination is substantially reduced.

5.5 Situating Curriculum Recall in Bloom's Taxonomy

Our benchmark explicitly targets the *Remember* level of Bloom's Revised Taxonomy [16], the lowest cognitive level, involving recognition and recall of specific facts, terminology, and structural details. This is a deliberate design choice: recall of factual curriculum content is a necessary precondition for higher-order tasks but is far from sufficient for effective AI-assisted teaching.

Huber and Niklaus [18] demonstrated that existing LLM benchmarks are heavily skewed toward lower-order cognitive skills. Our work contributes to this body of evidence by providing a granular evaluation of recall performance on a specific, practically important domain. However, we emphasise that high recall accuracy does not imply that a model can competently perform curriculum-related tasks that require higher-order thinking.

The gap between recall and pedagogical utility can be illustrated concretely. A model that correctly identifies AC9M3A01 as a Year 3 Number and Algebra outcome (Remember) may still be unable to explain why this concept is taught at Year 3 rather than Year 2 (Understand), generate an appropriate assessment task for this outcome (Apply), identify connections between this outcome and related Science outcomes (Analyse), judge whether a given lesson plan adequately addresses this outcome (Evaluate), or design a differentiated learning sequence that builds toward this outcome across terms (Create).

5.6 Limitations

Several limitations should be noted. First, our evaluation uses Gemini Flash 3 as the automated judge, which is also one of the models under test. While human validation confirmed approximately 80% judge accuracy and the deterministic fast-path checks resolve the majority of unambiguous cases without invoking the judge, this dual role may introduce subtle biases in borderline evaluations of Gemini 3 Flash's own responses. The 80% judge accuracy rate, while consistent with expected LLM-as-judge reliability, means that approximately one in five verdicts may be incorrect, a limitation that applies symmetrically across all models tested. Future work could employ multiple judges to quantify inter-judge agreement. Second, all curricula under test are currently undergoing multi-year revision processes, the Australian Curriculum transitioned from Version 8 to Version 9, and New Zealand is developing Te Mātaiaho, meaning that LLMs trained on earlier versions may produce answers that were once correct but are now outdated. Parametric knowledge of superseded curriculum content may reduce rather than improve accuracy. Third, sample sizes vary across curricula due to structural differences: the New Zealand Curriculum has 300 questions across only two categories, and the Western Australian Curriculum lacks content recall questions entirely. This means cross-curriculum comparisons should be interpreted with appropriate caution. Fourth, the benchmark tests only English-language curriculum content; it does not evaluate knowledge of Te Marautanga o Aotearoa (the Māori-medium NZ curriculum) or other language-specific frameworks. Fifth, while the methodology is described in sufficient detail to enable independent replication, the underlying structured curriculum database is not released due to copyright restrictions on state curriculum content and the proprietary nature of the database. Researchers wishing to reproduce this work must reconstruct structured representations from publicly available curriculum documents, which may introduce variation in parsing and normalisation. Finally, the benchmark evaluates curriculum knowledge at a single point in time; given the ongoing nature of curriculum reform across all jurisdictions tested, results should be interpreted as a snapshot rather than a permanent characterisation of model capabilities.

6. Future Work: Beyond Recall

The present benchmark establishes a foundation for evaluating LLM curriculum knowledge, but recall is only the first step. In this section, we outline a research agenda for extending evaluation to higher levels of Bloom's Revised Taxonomy [16], moving from testing what models remember to testing what they can do with curriculum knowledge.

6.1 Understand: Curriculum Comprehension

Tasks at the *Understand* level would assess whether models can interpret, explain, and contextualise curriculum content. Example benchmark tasks include: explaining the rationale for teaching a particular concept at a specific year level; summarising the learning progression for a concept across multiple year levels (e.g., how the treatment of fractions evolves from Year 1 to Year 7); classifying a set of outcomes by strand or sub-strand when given only their descriptions; and interpreting the relationship between a content descriptor and its associated elaborations.

These tasks require models to go beyond verbatim recall and demonstrate comprehension of curriculum structure and intent. Evaluation would require rubric-based judging with expert validation, since correct answers are less clearly defined than at the recall level.

6.2 Apply: Curriculum-Aligned Content Generation

Tasks at the *Apply* level would assess whether models can use curriculum knowledge to produce pedagogically appropriate outputs. Example tasks include: generating an assessment task (quiz, worksheet, or rubric) aligned to a specific content descriptor; creating a lesson activity targeting a specified outcome at the correct difficulty level for the year group; producing worked examples or model answers for a given outcome; and mapping a given resource or activity to the most relevant curriculum outcomes.

Evaluation at this level is substantially more complex, requiring expert teacher judges to assess both curriculum alignment and pedagogical quality. A hybrid evaluation approach combining automated rubric checks (e.g., does the generated task reference the correct outcome?) with human expert review of pedagogical appropriateness may be necessary.

6.3 Analyse: Cross-Curricular Connections

Tasks at the *Analyse* level would assess whether models can identify relationships, patterns, and connections within and across curricula. Example tasks include: identifying outcomes across different subjects that could be taught together in an integrated unit; comparing how a topic (e.g., sustainability) is treated across different subjects and year levels; analysing the prerequisite knowledge required for a given outcome; and detecting gaps or redundancies in a proposed scope and sequence.

These tasks are particularly relevant for Australian teachers who must identify cross-curricular integration opportunities.

6.4 Evaluate: Curriculum Alignment Judgement

Tasks at the *Evaluate* level would assess whether models can make informed judgements about curriculum alignment. Example tasks include: reviewing a lesson plan and identifying which curriculum outcomes it addresses (and which it claims but does not adequately address); critiquing a set of assessment tasks for alignment with stated learning intentions; evaluating whether a textbook chapter adequately covers the content descriptors for a given subject and year level; and judging the appropriateness of a resource for a specific year group.

This level of evaluation closely mirrors the expert judgements that curriculum coordinators and instructional leaders make daily. Benchmark design would require curated sets of lesson plans, assessment tasks, and resources with expert annotations of alignment quality.

6.5 Create: Curriculum Design and Planning

Tasks at the *Create* level, the highest in Bloom's hierarchy, would assess whether models can synthesise curriculum knowledge into novel, coherent outputs. Example tasks include: designing a term-long scope and sequence for a given subject and year level; creating a differentiated unit plan that addresses specified outcomes for diverse learners; producing a whole-school curriculum map showing how general capabilities are developed across year

levels; and designing formative and summative assessment strategies for a learning sequence.

Evaluation at this level would require substantial expert involvement, likely involving practising teachers reviewing generated plans against professional standards. This represents the most challenging and practically valuable extension of the current benchmark.

6.6 Methodological Considerations

Moving up Bloom's Taxonomy introduces significant methodological challenges. First, higher-order tasks have less clearly defined correct answers, requiring more sophisticated evaluation rubrics and likely human expert involvement. Second, task design must control for the influence of general intelligence versus specific curriculum knowledge, a model might produce a reasonable lesson plan through general pedagogical knowledge even without accurate curriculum recall. Third, evaluation at higher levels is more expensive and less scalable, necessitating smaller but more carefully designed benchmark sets. Fourth, the relationship between recall performance and higher-order performance is an empirical question: strong recall may be necessary but not sufficient, and the correlation between Bloom's levels merits investigation [18].

We suggest a phased approach: extending first to Understand (which can still be partially automated), then to Apply and Analyse (requiring expert validation of generated rubrics), and finally to Evaluate and Create (requiring substantial expert participation in both task design and evaluation).

7. Conclusion

We have presented the first systematic benchmark of LLM curriculum knowledge for Australian and New Zealand K-10 frameworks, evaluating seven frontier LLMs and one retrieval-augmented system across approximately 13,500 question-response pairs spanning four curricula and five categories of curriculum knowledge.

Our results reveal a striking knowledge deficit: no baseline LLM exceeds 41% overall pass rate across all curricula, with code-specific knowledge (code-to-description, content recall) near zero for most models, two of seven scored 0.0% on code-to-description. This is not a failure of model capability but of training data representation: the same models that achieve 95% on US standardised tests cannot recall the basic building blocks of Australian and New Zealand curriculum documents. The long-tail knowledge hypothesis [32] is strongly supported, with cross-curriculum analysis confirming that state-specific curricula (Victorian, Western Australian) are even less well represented than the national Australian Curriculum.

Retrieval augmentation, as demonstrated by CurricuLLM (88.7% overall), closes the gap dramatically, outperforming the best baseline by 47.7 percentage points and achieving 98.4% on open-ended topical questions. CurricuLLM's advantage is most pronounced on precisely the structured queries most relevant to teacher workflows: code-to-description (+76.5 pp) and description-to-code (+53.1 pp).

The error analysis reveals that outdated curriculum content is a persistent challenge, compounded by the fact that all tested curricula are currently undergoing multi-year revision

processes. This makes retrieval over up-to-date, authoritative curriculum databases not merely beneficial but essential for any AI tool supporting curriculum-aligned teaching.

Our benchmark contributes the first systematic evaluation of LLM curriculum knowledge for Australian and New Zealand K-10 frameworks. By explicitly grounding our work in Bloom's Taxonomy, we provide both a useful baseline and a roadmap for more ambitious evaluation of the pedagogical capabilities that teachers require from AI assistants. As AI tools become increasingly embedded in educational practice, rigorous, curriculum-specific benchmarks will be essential for ensuring these tools provide accurate, reliable, and educationally sound support.

References

[1] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. ICLR 2021. arXiv:2009.03300.

[2] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457.

[3] Welbl, J., Liu, N. F., & Gardner, M. (2017). Crowdsourcing Multiple Choice Science Questions. W-NUT Workshop at EMNLP 2017, pp. 94-106.

[4] Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W., & Duan, N. (2024). AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. Findings of NAACL 2024, pp. 2299-2314. arXiv:2304.06364.

[5] OpenAI. (2024). GPT-4 Technical Report. arXiv:2303.08774.

[6] Rodrigues, L., Pereira, F. D., Cabral, L., Ramalho, G., Gasevic, D., & Mello, R. F. (2024). Can GPT4 Answer Educational Tests? Empirical Analysis of Answer Quality Based on Question Complexity and Difficulty. AIED 2024, LNCS vol. 14829, Springer.

[7] Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track. arXiv:2306.05685.

[8] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. EMNLP 2023, pp. 2511-2522. arXiv:2303.16634.

[9] Wang, P., Li, L., Chen, L., et al. (2023). Large Language Models are not Fair Evaluators. arXiv:2305.17926.

[10] Li, J., et al. (2024). A Survey on LLM-as-a-Judge. arXiv:2411.15594.

[11] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020, 33, pp. 9459-9474. arXiv:2005.11401.

[12] Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-Augmented Generation for Educational Application: A Systematic Survey. Computers & Education: Artificial Intelligence. DOI: 10.1016/j.caeai.2025.100578.

[13] Dong, C. (2023). How to Build an Adaptive AI Tutor for Any Course Using Knowledge Graph-Enhanced Retrieval-Augmented Generation (KG-RAG). arXiv:2311.17696.

[14] Han, Z. F., Lin, J., Thomas, D. R., et al. (2024). Improving Assessment of Tutoring Practices using Retrieval-Augmented Generation. arXiv:2402.14594.

[15] Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain. New York: David McKay Company.

[16] Anderson, L. W., & Krathwohl, D. R. (Eds.) (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives (Complete Edition). New York: Longman.

[17] Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice, 41(4), 212-218.

[18] Huber, T., & Niklaus, C. (2025). LLMs meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations. COLING 2025, pp. 5211-5246.

[19] Kasneci, E., Seßler, K., Küchemann, S., et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 103, 102274.

[20] Australian Curriculum, Assessment and Reporting Authority [ACARA]. (2022). The Australian Curriculum, Version 9.0. https://v9.australiancurriculum.edu.au/

[21] Australian Government Department of Education. (2023). Australian Framework for Generative Artificial Intelligence (AI) in Schools. Canberra: Commonwealth of Australia.

[22] New Zealand Ministry of Education. (2007). The New Zealand Curriculum. Wellington. https://nzcurriculum.tki.org.nz/

[23] New Zealand Ministry of Education. (2024). Generative AI: Guidance and Resources for Education Professionals. https://www.education.govt.nz/school/digital-technology/generative-ai

[24] Coblenz, D., Dong, J., & Gibbs, B. (2025). Generative Artificial Intelligence in Aotearoa New Zealand Primary Schools: Teacher and Student Survey Findings. Wellington: NZCER.

[25] Bower, M., et al. (2025). What Generative Artificial Intelligence Priorities and Challenges Do Senior Australian Educational Policy Makers Identify (and Why)? The Australian Educational Researcher. DOI: 10.1007/s13384-025-00801-z.

[26] Imperial, J. M., Forey, G., & Tayyar Madabushi, H. (2024). Standardize: Aligning Language Models with Expert-Defined Standards for Content Generation. EMNLP 2024. arXiv:2402.12593.

[27] Liu, Z., Yin, S. X., Goh, D. H., & Chen, N. F. (2025). COGENT: A Curriculum-oriented Framework for Generating Grade-appropriate Educational Content. arXiv:2506.09367. BEA 2025 Workshop.

[29] Henkel, O., Boxer, A., Hills, L., & Roberts, B. (2024). Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability to Mark Short Answer Questions in K-12 Education. arXiv:2405.02985.

[30] OpenLearnLM Benchmark authors. (2025). OpenLearnLM Benchmark: A Unified Framework for Evaluating Knowledge, Skill, and Attitude in Educational Large Language Models. arXiv:2601.13882.

[31] Lelièvre, M., et al. (2025). Benchmarking the Pedagogical Knowledge of Large Language Models. arXiv:2506.18710.

[32] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. ACL 2023, pp. 9802-9822. arXiv:2212.10511.

[33] Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2024). Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. ICLR 2024 (Spotlight). arXiv:2305.13300.

[34] Xu, R., Qi, Z., Wang, C., Wang, H., Zhang, Y., & Xu, W. (2024). Knowledge Conflicts for LLMs: A Survey. EMNLP 2024. arXiv:2403.08319.

[35] Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-Based Knowledge Conflicts in Question Answering. EMNLP 2021, pp. 7052-7063.

[36] Liang, P., Bommasani, R., Lee, T., et al. (2023). Holistic Evaluation of Language Models. Transactions on Machine Learning Research (TMLR). arXiv:2211.09110.

[37] Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). Benchmark Data Contamination of Large Language Models: A Survey. arXiv:2406.04244.

[38] Deng, C., Zhao, Y., Tang, X., Gerstein, M., & Cohan, A. (2024). Investigating Data Contamination in Modern Benchmarks for Large Language Models. NAACL 2024, pp. 8706-8719.

[39] White, C., Dooley, S., Roberts, M., Pal, A., et al. (2025). LiveBench: A Challenging, Contamination-Limited LLM Benchmark. ICLR 2025 (Spotlight). arXiv:2406.19314.

[40] Chang, Y., Wang, X., Wang, J., et al. (2024). A Survey on Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 15(3), Article 39, pp. 1-45. arXiv:2307.03109.