From Recall to Reasoning: Evaluating Cognitive Depth Annotation for Educational Content

Anonymous ACL submission

Abstract

Large language models are increasingly used in educational settings, yet most existing evaluations focus primarily on problem-solving accuracy rather than the ability to organize educational texts according to pedagogical objectives. This paper introduces DepthEdu, a benchmark and distillation pipeline for evaluating language models as cognitive organizers of educational text. DepthEdu requires models to assign structured pedagogical labels across three dimensions: Bloom's cognitive process, Bloom's knowledge dimension, and Webb's Depth of Knowledge. The dataset contains 7,330 training examples and 2,000 test examples drawn from FineWeb-Edu, MATH, GPQA, GSM8K, and MATH500, with reference labels generated by Qwen3-32B. We evaluate a range of small language models using label-level accuracy, exact match, DOK macro-F1, and per-example annotation latency. Results show that off-the-shelf small models struggle particularly with DOK classification, highlighting cognitive complexity as a key bottleneck. To improve efficiency and accuracy, we fine-tune Qwen3-0.6B with supervised learning, producing OB-Distill-0.6B. The distilled model substantially outperforms its base model, achieving SOTA accuracy while maintaining fast annotation speed.

1 Introduction

With the rapid development of large language models, selecting appropriate questions and data has become increasingly important for pre-training, reinforcement learning, and capability evaluation. Educationally oriented data are now drawn from a heterogeneous mixture of learning-oriented web text, mathematical problem sets, scientific questions, and benchmark-style assessment items. These data are often treated as input-output instances: a model is trained or evaluated by whether it can predict the next tokens, solve the problem, or produce the correct answer. However, educational content also carries cognitive structure. A short passage may support factual recall, a mathematics problem may require routine procedural application, and a science question may demand strategic reasoning across multiple concepts. Understanding the cognitive depth of such content is therefore important for characterizing what kinds of learning objectives and reasoning demands are represented in the data used to train and evaluate educational language models.

The importance of cognitive depth has long been studied in education. In curriculum and assessment research, educational materials are often evaluated not only by topic coverage or correctness, but also by the forms of knowledge and thinking they elicit. Bloom's Taxonomy (Bloom et al., 1956; Darwazeh and Branch, 2015) provides one of the most influential frameworks for this purpose, and its revised version separates educational objectives into two complementary dimensions: the cognitive process dimension and the knowledge dimension. The cognitive process dimension distinguishes remembering, understanding, applying, analyzing, evaluating, and creating, while the knowledge dimension distinguishes factual, conceptual, procedural, and metacognitive knowledge. Webb's Depth of Knowledge (Webb, 1997a) further complements Bloom's Taxonomy by focusing on the complexity of reasoning required by a task. Motivated by this multidimensional view, we formulate cognitive depth annotation as a structured prediction task over three pedagogical axes: Bloom Cognitive Process, Bloom Knowledge Type, and Webb's Depth of Knowledge.

We introduce DepthEdu, a benchmark and distillation pipeline that uses a high-capacity teacher annotator and evaluates whether smaller models can reproduce its pedagogical judgments efficiently. The current release contains 7,330 training examples and 2,000 test examples from FineWeb-Edu (Lozhkov et al., 2024), MATH (Hendrycks et al.,

2021), GPQA (Rein et al., 2023), GSM8K (Cobbe et al., 2021), and MATH500. We evaluate a range of small models and report both label accuracy and annotation latency. We also train Qwen3-0.6B with full-parameter supervised fine-tuning (SFT), producing ten checkpoints across epochs.

Our contributions are:

1. We construct a dataset of 7,330 training examples and 2,000 test examples from five educational sources, covering both problem statements and educational prose.

2. We introduce DepthEdu, a benchmark for evaluating off-the-shelf small language models on label accuracy, exact match, DOK macro-F1, and annotation latency.

3. We train ED-Distill-0.6B, an efficient student model that provides strong performance and a scalable starting point for large-scale pedagogical annotation.

2 Related Work

Educational LLM evaluation. LLMs are now studied across a broad range of educational applications, including tutoring (Wei et al., 2025), question generation (Elkins et al., 2024) and other educational scenarios. Much of the existing benchmarking literature still treats educational ability primarily as problem-solving accuracy (Cobbe et al., 2021; Hendrycks et al., 2021; Rein et al., 2023). These benchmarks are important for measuring subject-matter reasoning, but they do not directly evaluate whether a model can identify the pedagogical role of an educational text.

Learning-science taxonomies. Bloom's taxonomy (Bloom et al., 1956) and its revised version (Anderson and Krathwohl, 2001) remain foundational frameworks for categorizing educational objectives. Depth of Knowledge (DOK) (Webb, 1997b) complements Bloom's taxonomy by shifting attention from the surface difficulty of a task to the cognitive complexity required to complete it. Building on these learning-science frameworks, recent studies have begun to operationalize educational taxonomies within NLP systems. For example, BloomNet (Waheed et al., 2021) classifies learning outcomes according to Bloom levels, while Mohammed and Omar (2020) employed Word2Vec to automatically classify questions into the six cognitive levels of the original Bloom's taxonomy. Beyond classification, other work has used LLMs to generate quiz questions aligned with Bloom's objectives (Elkins et al., 2024) or to compare generation strategies across different Bloom skill levels (Scaria et al., 2024). In parallel, Essential-Web (AI et al., 2025) annotated 24T tokens of web data to train a 0.5B model. However, these approaches either focus primarily on question generation, address only a single pedagogical dimension, or rely on large-scale corpora that are not specifically educational.

3 DepthEdu

3.1 Task Definition

Given an input educational text x, the model $\mathcal{M}$ predicts a structured tuple,

$$\mathcal{M}(x) = y^{pred} = (y_c^{pred}, y_k^{pred}, y_d^{pred}). \quad (1)$$

Here y_c is the Bloom cognitive process (BTC) label, y_k is the Bloom knowledge dimension (BTK) label, y_d is the DOK label, and y^{pred} denotes the structured label tuple predicted by $\mathcal{M}$. We evaluate each field separately and also report exact match (EM), which requires all three labels to match the reference tuple. This can be formalized as follows:

$$\mathrm{EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(y_i^{pred} = y_i\right), \quad (2)$$

where $\mathbf{1}(\cdot)$ is the indicator function and y is the ground truth.

3.2 Data Sources

DepthEdu combines random examples from five sources: FineWeb-Edu, MATH, GPQA, GSM8K, and MATH500. FineWeb-Edu contributes educational prose, while the remaining datasets contribute math and science problem statements. Figure 1 reports the split composition.

The test split contains 2,000 examples and the training split contains 7,330 examples. The test set is constructed to include substantial coverage from each source, with 800 FineWeb-Edu examples and 300 examples from each of the four problem datasets. The training set contains more FineWeb-Edu and MATH examples, reflecting the available local data mixture.

3.3 Teacher Annotation

We use Qwen3-32B as the teacher annotator. The teacher receives definitions for BTC, BTK, and

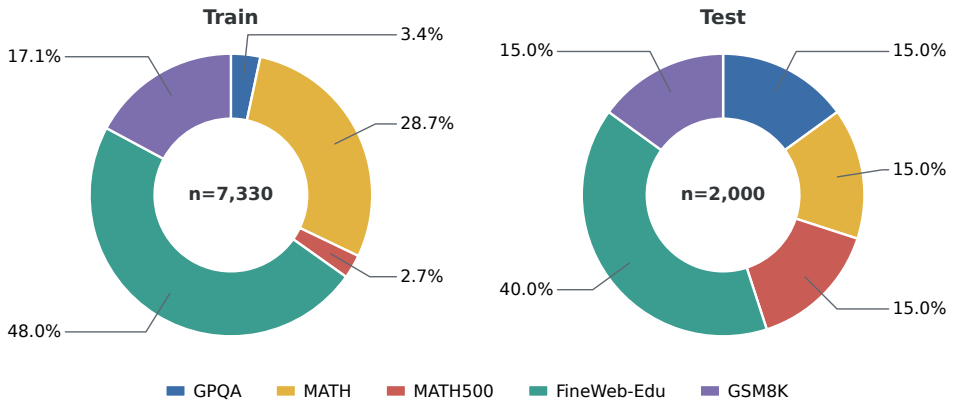

Figure 1: Percentage composition of the Qwen3-32B-labeled DepthEdu splits. Percentages are rounded to one decimal place.

Dim.	Label	Train	Test
BTC	Analyze	301	130
BTC	Apply	3847	1117
BTC	Create	62	18
BTC	Evaluate	106	35
BTC	Remember	604	153
BTC	Understand	2410	547
BTK	Conceptual	2861	780
BTK	Factual	1092	279
BTK	Metacognitive	1	0
BTK	Procedural	3376	941
DOK	Extended thinking	2	0
DOK	Recall and Reproduction	888	218
DOK	Skill/Concept	2742	678
DOK	Strategic Thinking and Reasoning	3698	1104

Table 1: Label distribution under Qwen3-32B supervision. BTC, BTK, and DOK denote Bloom cognitive process, Bloom knowledge type, and Depth of Knowledge.

DOK, and returns a structured JSON object with one label for each field. The resulting teacher outputs are used as references for evaluation and as supervision for SFT. Table 1 shows the resulting label distribution. Additional details are provided in Appendix A.

3.4 Training Data Conversion

For supervised training, each teacher-labeled example is converted into a chat-format record. The user message contains the educational text, and the assistant message is the JSON label object containing the three evaluated fields. Method to process long documents that exceed the context budget is provided in Appendix C.

4 Experiments

Evaluated models. We evaluate off-the-shelf small models from the Qwen and Llama families, including Qwen3.5-4B, Qwen3-4B-Instruct-2507, Qwen3-4B (Yang et al., 2025), Llama-3.2-3B-Instruct, Qwen3.5-2B, Qwen3.5-0.8B, Qwen2.5-1.5B-Instruct (Qwen et al., 2025), Qwen2.5-1.5B, Qwen3-0.6B, Qwen3-1.7B, and EAI-distill-0.5B (AI et al., 2025). In addition, we evaluate ten full-parameter SFT checkpoints of Qwen3-0.6B, each saved after one training epoch. The results are reported in Table 2. We select the epoch-10 checkpoint as ED-Distill-0.6B.

Metrics. We report accuracy for each label dimension, average accuracy across the three dimensions, and per-example annotation latency in Table 2. Additional metrics, including exact match over the complete three-label tuple, balanced accuracy, macro precision, macro recall, macro F1, and weighted F1, are reported in Appendix B. The latency metric is measured by the labeling pipeline as total model response time divided by the number of evaluated examples.

Implementation details. All experiments are conducted on NVIDIA RTX 4090 GPUs, with inference performed using SGLang (Zheng et al., 2024). For SFT, we use Qwen3-0.6B as the base model and train it on the 7,330 examples labeled by Qwen3-32B with swift (Zhao et al., 2024). We perform full-parameter SFT for 10 epochs using a learning rate of 1×10^{-5}, a per-device batch size of 8, a warmup ratio of 0.05, DeepSpeed ZeRO-3, the Liger kernel, and epoch-level checkpointing. The checkpoint from each epoch is evaluated on the 2,000-example test split. During teacher annotation, we enable thinking for Qwen3-32B.

4.1 Results

Table 2 shows that off-the-shelf small models vary substantially. EAI-Distill-0.5B obtains 77.3% Bloom cognitive accuracy, 69.5% Bloom knowl-

Model	Bloom Cognitive Process	Bloom Knowledge Type	Webb's Depth of Knowledge	Avg. Time (s)	Avg. Accuracy
Qwen3.5-4B	81.0%	77.6%	41.1%	1.305	66.6%
Qwen3-4B-Instruct-2507	70.6%	75.3%	57.0%	0.702	67.7%
EAI-Distill-0.5B	77.3%	69.5%	38.1%	0.540	61.6%
Qwen3-4B	57.6%	68.0%	50.6%	0.709	58.8%
Llama-3.2-3B-Instruct	70.0%	39.5%	50.5%	0.880	53.3%
Qwen3.5-2B	44.5%	25.6%	25.0%	0.973	31.7%
Qwen3.5-0.8B	61.2%	24.2%	13.1%	1.159	32.8%
Qwen2.5-1.5B-Instruct	64.9%	27.7%	15.5%	0.492	36.0%
Qwen2.5-1.5B	55.6%	23.9%	16.0%	**0.477**	31.8%
Qwen3-0.6B	32.4%	39.0%	12.4%	0.527	27.9%
Qwen3-1.7B	30.8%	14.6%	30.6%	0.528	25.3%
OB-Distill-0.6B	**86.2%**	**78.8%**	**73.6%**	0.630	**79.5%**

Table 2: Small-model evaluation against Qwen3-32B labels on the 2,000-example test split. Avg. Accuracy is the mean accuracy across BTC, BTK, and DOK. Average time is per-example labeling latency recorded by the evaluation pipeline. More metrics can be found in Appendix B.

edge accuracy, 38.1% DOK accuracy, and 61.6% average accuracy on our 2,000-example test split. Among the other evaluated off-the-shelf models, Qwen3.5-4B achieves strong Bloom cognitive and knowledge accuracy but lower DOK accuracy, yielding 66.6% average accuracy. Qwen3-4B-Instruct-2507 obtains the best off-the-shelf average accuracy at 67.7%. Several smaller models collapse on at least one dimension: Qwen3-0.6B reaches only 27.9% average accuracy, and Qwen3-1.7B obtains 25.3%.

The results also show that speed and accuracy are not aligned. Qwen2.5-1.5B is fast at 0.477 seconds per example but has low average accuracy. Qwen3.5-4B is slower at 1.305 seconds per example but much more accurate.

4.2 Analysis

DOK is the bottleneck. Across off-the-shelf models, DOK accuracy is often much lower than Bloom cognitive or knowledge accuracy. This is expected because DOK requires distinguishing cognitive complexity from difficulty or domain sophistication. A model may recognize a task as procedural while still misjudging whether it requires routine skill use or strategic reasoning.

Fast annotation becomes feasible. The teacher labels are expensive: the stored Qwen3-32B test annotations average 9.31 seconds per example, while the training annotations average 31.01 seconds per example. In contrast, the best SFT checkpoint annotates the test set at 0.647 seconds per example, a 14.4× speedup over the stored teacher test annotations in our pipeline. Essential-Web reports a larger 50× throughput gain by combining model shrinking, context distillation, and generation condensation. Our current system does not yet use all of those optimizations, so the main result is quality transfer into a small model rather than maximal serving throughput.

5 Conclusion

We introduced DepthEdu, a benchmark and distillation pipeline for evaluating whether language models can annotate educational content according to structured pedagogical taxonomies. DepthEdu represents each item through Bloom's cognitive process, Bloom's knowledge dimension, and Webb's Depth of Knowledge.

Our experiments show that this annotation problem remains challenging for off-the-shelf small models. While several models achieve reasonable performance on Bloom-style labels, DOK classification is consistently harder, suggesting that reasoning depth is less directly captured by general instruction-following ability. Full-parameter SFT of Qwen3-0.6B substantially improves all three dimensions, especially DOK, and yields a compact annotator with much lower latency than the teacher pipeline.

These results suggest that teacher-generated pedagogical supervision can transfer useful cognitive annotation behavior into small models, making large-scale analysis of educational corpora more practical. Future work should incorporate expert human validation, improve coverage of rare taxonomic categories, and extend the annotation schema beyond the three fields studied here.

Limitations

DepthEdu currently has several limitations. The reference labels are produced by a model but expert human annotators. The benchmark therefore evaluates agreement with a strong teacher model rather than pedagogical correctness in an absolute sense. Second, label imbalance is severe for rare categories such as METACOGNITIVE and EXTENDED THINKING, limiting claims about those classes. Third, the current task evaluates only three taxonomic fields; the repository contains placeholders for disciplinary classification, but those fields are not active in the current evaluation. Fourth, the dataset mixes educational prose and problem statements from existing sources, so licensing and representativeness should be reviewed before public release. Finally, latency numbers depend on the local serving stack and hardware, and should be interpreted as relative within this experimental setup.

References

Essential AI, :, Andrew Hojel, Michael Pust, Tim Romanski, Yash Vanjani, Ritvik Kapila, Mohit Parmar, Adarsh Chaluvaraju, Alok Tripathy, Anil Thomas, Ashish Tanwer, Darsh J Shah, Ishaan Shah, Karl Stratos, Khoi Nguyen, Kurt Smith, Michael Callahan, Peter Rushton, and 6 others. 2025. Essential-web v1.0: 24t tokens of organized web data. *arXiv preprint arXiv:2506.14111*.

Lorin W Anderson and David R Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, and 1 others. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Afnan N Darwazeh and Robert Maribe Branch. 2015. A revision to the revised bloom's taxonomy. *2015 Annual Proceedings–Indianapolis*, 2:220–225.

Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. 2024. How teachers can use large language models and bloom's taxonomy to create educational quizzes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):23084–23091.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. Fineweb-edu: the finest collection of educational content. *arXiv preprint arXiv:2406.17557*.

Manal Mohammed and Nazlia Omar. 2020. Question classification based on bloom's taxonomy cognitive domain using modified tf-idf and word2vec. *PloS one*, 15(3):e0230442.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. *arXiv preprint arXiv:2408.04394*.

Abdul Waheed, Muskan Goyal, Nimisha Mittal, Deepak Gupta, Ashish Khanna, and Moolchand Sharma. 2021. Bloomnet: A robust transformer based model for bloom's learning outcome classification. *Proceedings of the 4th International Conference on Natural Language and Speech Processing*, pages 209–218.

Norman Webb. 1997a. Research monograph number 6: Criteria for alignment of expectations and assessments on mathematics and science education. *Washington, DC: CCSSO*.

Norman L Webb. 1997b. Criteria for alignment of expectations and assessments in mathematics and science education. research monograph no. 6. *Education Resources Information Center*.

Shouang Wei, Min Zhang, Xin Lin, Bo Jiang, Kun Kuang, and Zhongxiang Dai. 2025. Uco: A multi-turn interactive reinforcement learning method for adaptive teaching with large language models. *arXiv preprint arXiv:2511.08873*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *arXiv preprint arXiv:2408.05517*.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody H Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, and 1 others. 2024. Sglang: Efficient execution of structured language model programs. *Advances in neural information processing systems*, 37:62557–62583.

A Prompt Templates and Annotation Instructions

This section provides the prompt templates and annotation guidelines used to obtain and evaluate the taxonomy labels in OrganizerBench. Each text is annotated along three educational dimensions: Bloom's cognitive process dimension (BTC), Bloom's knowledge dimension (BTK), and Depth of Knowledge (DOK). The annotator is instructed to choose exactly one label for each dimension and to base the decision only on the given text.

A.1 Label Space

Bloom's cognitive process dimension. The allowed BTC labels are `Remember`, `Understand`, `Apply`, `Analyze`, `Evaluate`, and `Create`. These labels correspond respectively to recalling information, explaining or interpreting meaning, using a method or concept, analyzing structure or evidence, making judgments based on criteria, and generating or designing new outputs.

Bloom's knowledge dimension. The allowed BTK labels are `Factual`, `Conceptual`, `Procedural`, and `Metacognitive`. Factual knowledge covers terminology and isolated details; conceptual knowledge covers categories, principles, systems, and relationships; procedural knowledge covers methods, steps, techniques, and workflows; and metacognitive knowledge covers learning strategies, self-monitoring, reflection, planning, and regulation.

Depth of Knowledge. The allowed DOK labels are `Recall and Reproduction`, `Skill/Concept`, `Strategic Thinking and Reasoning`, and `Extended thinking`. DOK is judged by cognitive complexity rather than text length, subject difficulty, or surface sophistication.

A.2 Annotation Instructions

The annotator follows four main principles. First, labels must be assigned based only on the text itself, without relying on external knowledge unless it is directly supported by the text. Second, when a text contains multiple actions or topics, the annotator chooses the label corresponding to the central learning goal rather than a secondary action. Third, for explanatory texts, the label reflects the primary learning goal supported by the explanation. Fourth, the output must contain only the structured label fields, with values exactly matching the allowed label names.

A.3 Teacher Annotation Prompt

For teacher-model annotation, we use a system instruction that defines the role, label definitions, and decision rules, followed by a user prompt containing the target text. The response is constrained to a JSON schema with the three taxonomy fields. To keep the appendix compatible with a two-column layout, we summarize the prompt in wrapped text rather than fixed-width verbatim blocks.

System instruction. The annotator is told: "You are a strict educational text annotator. Your task is to assign the single best label for the given text based only on the text itself." The instruction further requires the annotator to avoid outside knowledge unless it is directly supported by the text, choose the most central label when multiple actions or topics appear, output only the structured result required by the schema, and use label values that exactly match the allowed enum values.

The system instruction also includes the full label definitions. For BTC, `Remember` covers identifying, recalling, listing, recognizing, defining, and restating; `Understand` covers explaining, summarizing, comparing, classifying, and interpreting meaning; `Apply` covers using a rule, formula, method, or concept to complete a task; `Analyze` covers breaking down structure, identifying relationships, analyzing causes, and examining evidence; `Evaluate` covers judging, critiquing, justifying, defending, or choosing based on criteria; and `Create` covers designing, generating, proposing, composing, or producing something new.

For BTK, `Factual` covers terminology, details, isolated facts, and definitions; `Conceptual` covers principles, models, categories, systems, and relationships; `Procedural` covers steps, algorithms, methods, techniques, workflows, and solution

procedures; and Metacognitive covers learning strategies, self-monitoring, reflection, regulation, and planning.

For DOK, Recall and Reproduction covers direct recall or routine reproduction; Skill/Concept covers basic understanding, standard application, and familiar procedures; Strategic Thinking and Reasoning covers evidence-based reasoning, comparison, multi-step thinking, and strategy selection; and Extended thinking covers multi-stage synthesis, project-like work, long-horizon inquiry, and open-ended integration. The annotator is explicitly instructed to judge DOK by cognitive complexity rather than by difficulty, length, or subject sophistication.

User instruction. The user prompt asks the model to read the text and assign exactly one label for each of the following fields: bloom_taxonomy_cognitive, bloom_taxonomy_knowledge, and depth_of_knowledge. The prompt further instructs the model to prioritize the main cognitive task required of the learner rather than minor or secondary actions, and to label explanatory texts according to the primary learning goal they support. The target document is inserted between <text> and </text> tags.

The expected JSON output contains only the three fields bloom_taxonomy_cognitive, bloom_taxonomy_knowledge, and depth_of_knowledge.

A.4 SFT Model Prompt Template

For supervised fine-tuning and SFT model evaluation, the model receives a compact chat template. The system message is a task identifier, the user message contains the input text, and the assistant message is the JSON label object. Long documents are truncated with a beginning–middle–end strategy before being inserted into the user message.

The chat format is: system message Organizerbench; user message containing the possibly truncated input text; and assistant message containing a JSON object with bloom_taxonomy_cognitive, bloom_taxonomy_knowledge and depth_of_knowledge.

At inference time, the SFT model is prompted with the same system and user messages and is evaluated on the three taxonomy fields.

B Additional Evaluation Metrics

Table 3 provides complementary aggregate metrics beyond the main per-dimension accuracy, average accuracy, and latency results reported in the main text. Specifically, we report exact match, balanced accuracy, macro precision, macro recall, macro F1, and weighted F1. These metrics provide a more complete view of model behavior under label imbalance, especially for rare cognitive categories. ED-Distill-0.6B achieves the strongest overall performance across all additional metrics, reaching 53.5 exact match, 67.5 balanced accuracy, 72.8 macro precision, 67.5 macro recall, 69.5 macro F1, and 79.1 weighted F1. Compared with the strongest off-the-shelf baselines, the distilled model shows substantially better macro-level performance, indicating that supervised distillation improves not only frequent-label accuracy but also robustness across the broader cognitive-depth label space.

C Ultra Long Document Handle

For documents that exceed the annotation context budget, we apply a simple beginning–middle–end truncation strategy before converting them into chat-format records. The maximum input length is set to 30,000 characters. If a document is within this limit, it is kept unchanged. Otherwise, the budget is divided into three spans of 10,000 characters each: the first 10,000 characters from the beginning of the document, the last 10,000 characters from the end, and a continuous 10,000-character span sampled from the middle region. Specifically, we randomly select a center point from the valid middle range and retain the 5,000 characters before and after that point. This ensures that the sampled middle span does not overlap with the retained prefix or suffix.

The truncated input is formatted with explicit segment markers: [beginning], [middle], and [end]. This preserves the opening context, representative internal content, and closing context while keeping the input within the annotation budget. We use this strategy because long educational texts often introduce the topic at the beginning, develop the main learning content in the body, and summarize or conclude the intended task near the end.

D Representative Label Examples

This appendix provides concrete examples from the test split. We omit sample identifiers for readabil-

Model	Exact Match	Balanced Acc	Macro P	Macro R	Macro F1	Weighted F1
Qwen3.5-4B	30.9	<u>62.5</u>	<u>64.6</u>	<u>62.5</u>	<u>58.2</u>	63.8
Qwen3-4B-Instruct-2507	<u>34.2</u>	57.6	60.9	53.1	52.3	<u>68.2</u>
EAI-Distill-0.5B	23.7	50.1	55.5	50.1	46.9	59.3
Qwen3-4B	21.9	54.4	48.4	44.6	41.7	60.4
Llama-3.2-3B-Instruct	11.4	33.7	48.4	33.7	28.6	45.6
Qwen3.5-2B	9.3	40.1	40.6	36.5	24.6	32.0
Qwen3.5-0.8B	2.1	38.0	22.5	38.0	21.2	27.7
Qwen2.5-1.5B-Instruct	2.1	30.1	28.0	27.5	20.6	32.1
Qwen2.5-1.5B	0.9	24.8	19.8	22.7	16.0	26.4
Qwen3-0.6B	0.6	22.2	24.6	21.2	14.6	23.8
Qwen3-1.7B	0.7	30.3	33.0	30.3	17.6	26.3
ED-Distill-0.6B (*Ours*)	**53.5**	**67.5**	**72.8**	**67.5**	**69.5**	**79.1**

Table 3: Additional evaluation metrics against Qwen3-32B labels on the 2,000-example test split. Exact Match is the percentage of examples for which all three label dimensions match. Macro precision, macro recall, macro F1, and weighted F1 are computed per label dimension and averaged across the three dimensions. Bold and underlined values denote the best and second-best results for each metric, respectively.

ity, but report the source dataset and a shortened description of the input content.

D.1 Bloom's Cognitive Process Dimension

Remember. *FineWeb-Edu.* One passage gives factual information about mule deer: "The mule deer ... is an exclusively western species commonly seen in open-brush country ..." The main task is to recognize or recall descriptive facts. *FineWeb-Edu.* Another passage presents a short biography of Kurt Vonnegut, including his dates, occupation, and major work. This is labeled as recall-oriented factual recognition.

Understand. *FineWeb-Edu.* A passage explains three types of farming, including sustainable, commercial, and subsistence farming, and asks the reader to distinguish their meanings. *GPQA.* A chemistry question states that bromination is stereospecific but that a particular alkene gives both cis and trans products, then asks the model to rationalize the concept behind the observation.

Apply. *MATH.* One problem asks: "The four-digit numeral 3AA1 is divisible by 9. What digit does A represent?" Solving it requires applying the divisibility rule for 9. *MATH.* Another problem gives an algebraic condition involving $x^3 + 1/x^3$ and asks for $x + 1/x$, requiring application of a standard algebraic identity.

Analyze. *GPQA.* One case describes an experiment with Escherichia coli, ethidium bromide, and antibiotic inhibition zones. The learner must analyze how the experimental changes relate to antibiotic resistance. *GPQA.* Another case describes a yeast protein deletion mutant with reduced growth rate and cell size, followed by interactome evidence. The task requires reasoning about relationships among phenotype, interaction, and biological function.

Evaluate. *GPQA.* One case describes ethylene polymerization with a homogeneous organometallic catalyst and asks which catalyst strategy is appropriate. The learner must evaluate alternatives under chemical constraints. *GPQA.* Another case asks which statement about long non-coding RNAs is untrue, requiring judgment among competing claims.

Create. *FineWeb-Edu.* One classroom resource asks students to work with number lines, including tasks such as "create line using an equation" and "create equation using a line." *FineWeb-Edu.* Another art lesson asks students to design a Hogwarts-style house banner after studying symbolism, animals, and color choices in production design.

D.2 Bloom's Knowledge Dimension

Factual. *FineWeb-Edu.* The mule deer passage lists species-level facts, habitat information, and descriptive details. *FineWeb-Edu.* The Vonnegut passage provides biographical facts such as dates, authorship, and a major literary work.

Conceptual. *FineWeb-Edu.* The farming passage explains relationships among farming categories and how they differ from each other. *GPQA.* The bromination question centers on stereospecificity and reaction-mechanism concepts rather than isolated facts.

Procedural. *FineWeb-Edu.* A music-teaching passage gives a step-by-step procedure for teach-

ing a new piano piece, including finding patterns, rhythms, intervals, chord shapes, and accidentals. *MATH.* The divisibility problem requires following a numerical procedure to determine the unknown digit.

`Metacognitive.` No `Metacognitive` examples appear in the 2,000-example test split.

D.3 Depth of Knowledge

`Recall and Reproduction.` *FineWeb-Edu.* The mule deer passage mainly requires reproducing factual information from a descriptive text. *FineWeb-Edu.* The Vonnegut biography similarly supports direct recall of explicitly stated facts.

`Skill/Concept.` *GSM8K.* One arithmetic word problem compares two quantities of kangaroos and asks how many days are needed for one farmer to match the other's current count. *MATH.* One algebra problem asks the learner to simplify a difference between two products of linear expressions.

`Strategic Thinking and Reasoning.` *FineWeb-Edu.* The piano-teaching passage requires selecting and coordinating multiple strategies, such as identifying patterns, rhythms, intervals, chord shapes, and key signatures. *GPQA.* The antibiotic-resistance case requires interpreting an experimental setup and comparing evidence before and after exposure to ethidium bromide.

`Extended thinking.` No `Extended thinking` examples appear in the 2,000-example test split.