

Generating Curriculum Grounded Nepali Questions Based On Bloom’s Taxonomy

Prabigya Acharya

Abstract

The development of conversational large language models has brought a drastic change in educational settings, both in school and at home. While the Curriculum Development Centre of Nepal publishes detailed guides for teachers to create and use teaching materials in their classrooms, for subjects like Nepali these remain largely unused: theoretical topics are taught by rote learning and reproduced in examinations, making it a subject in which students score lowest and feel least confident. With the advent of LLMs, there is a real opportunity to turn these unused guidelines into large amounts of diverse teaching and evaluation material without adding to teachers’ burden. In this work, we use conversational LLMs to generate curriculum-grounded questions for the Grade 7 and 10 Nepali subject and evaluate their usability in a classroom setting. Results show that Nepali objectives require more than one Bloom’s taxonomy level to define them: forcing a single level per objective drops teacher agreement to near chance $\kappa = 0.13$, while allowing multiple levels recovers agreement on at least one level for 84.6% of objectives. We also observe that LLMs consistently dumb down the taxonomy level of objectives, plateauing at roughly 0.50 Macro F1, whereas a supervised classifier over frozen multilingual encoders reaches 0.612. Finally, teachers evaluated the generated questions through a modified eight-item rubric: grounding questions in the relevant taxonomy levels raised teacher endorsement from 12.5% to 88%. The question generation tool made from this approach is currently deployed in a government school in Nepal.

1 Introduction

Even though Nepali is a central subject in Nepal’s school curriculum, national assessment data point to a persistent problem in how it is taught. In the 2023 National Assessment of Student Achieve-

ment, Nepali recorded the lowest mean score (Education Review Office, 2025). Classroom practice is typically teacher-centred and rote memorisation is encouraged (Bhattarai and Basnet, 2022), aimed at reproducing a fixed set of examination questions. The Curriculum Development Centre publishes learning objectives to support objective-aligned teaching (Curriculum Development Centre, 2025, 2026), but authoring diverse questions aligned with every objective is a major manual burden, falling hardest on schools with the fewest teachers. As a result, the guidelines remain hardly used. Reducing this authoring cost is therefore not only a convenience but a contribution to educational equity in a public school system serving a low-resource language, in line with United Nations Sustainable Development Goal 4 on quality education.

Large language models (LLMs) offer a way to turn these unused curriculum guides into teaching material at scale. Prior works generate questions conditioned on Bloom’s taxonomy levels (Scaria et al., 2024b), that supports teachers to build quizzes (Elkins et al., 2024), and extends to non-Western curricula such as Indian high-school social science (Scaria et al., 2024a). No comparable study exists for Nepali curriculum, a low-resource language for which such pipelines remain untested. In this work we generate curriculum-grounded questions for the Grade 7 and Grade 10 Nepali subject and evaluate, with practising teachers, whether the resulting questions are usable in a real classroom.

Generating questions this way requires first assigning Bloom’s levels to the curriculum objectives that condition generation. Prior approaches assign a single level per objective (Kolagar et al., 2025; Scaria et al., 2024b), and Li et al. (2022) note that university course objectives can in fact demand several cognitive levels at once. We find that Nepali objectives, which combine factual comprehension with creative writing and literary analysis, are gen-

uniquely multi-dimensional: two teachers forced to assign one level per objective agreed near chance, whereas when they were allowed multiple labels, they agreed on at least one level for the large majority of objectives. Forcing a single label discards cognitive demand that the objective actually carries, and we accordingly generate questions under both single-label and multi-label taxonomy conditions.

Building this pipeline surfaces a second finding. Zero-shot LLM taxonomy extraction is unreliable in model-specific ways: across three model families, every model suppresses at least one Bloom’s category, but which category collapses differs by model rather than tracking model quality, indicating that the bias comes from each model’s internal category priors rather than from the objectives. A lightweight supervised classifier over frozen multilingual encoder representations avoids these blind spots and recovers the categories that zero-shot prompting misses. We then ground question generation in the resulting taxonomy labels and have Nepali-subject teachers evaluate the questions on an adapted bilingual pedagogical rubric. Our contributions are:

- A pipeline for curriculum-aligned question generation for the Grade 7 and Grade 10 Nepali subject, producing diverse questions grounded in per-objective learning goals under both single- and multi-label Bloom’s taxonomy conditions, with a teacher-based usability evaluation.
- A dataset of **93** Nepali learning objectives annotated with multi-label Bloom’s Revised Taxonomy levels by two practising teachers, together with **400** generated Nepali questions.
- Evidence across three model families that zero-shot taxonomy extraction carries model-specific category biases, and that a linear classifier over frozen multilingual encoders avoids them.
- A curriculum grounded question generator, currently being used in a government school, that we will open source.

2 Related Work

Mapping learning objectives to Bloom’s taxonomy. Aligning learning objectives with Bloom’s Revised Taxonomy (Anderson and Krathwohl,

2001) has been studied for quality assurance in education, initially with non-LLM methods for classifying questions and objectives, ranging from keyword dictionaries (Chang and Chung, 2009) and TF-IDF or SVM classifiers (Mohammed and Omar, 2020; Yahya et al., 2012) to transformer-based models (Waheed et al., 2021), which performed adequately on lower-order categories but degraded on higher-order ones. Li et al. (2022) introduced a large dataset of over 21,000 learning objectives and evaluated traditional and BERT-based classifiers, but treated each objective as carrying a single cognitive skill. Most relevant to our setting, Kolagar et al. (2025) compared LLM- and non-LLM-based methods for multi-label mapping of university course objectives, using expert annotations as the gold standard, and reported that LLMs outperform non-LLM baselines without exhibiting strong selection biases. Our findings diverge on both axes: in the Nepali school curriculum we observe pronounced, model-specific category suppression under zero-shot LLM extraction, and a supervised classifier over frozen multilingual encoders outperforms the zero-shot LLMs. This contrast suggests that conclusions drawn from English and German university objectives do not transfer directly to a low-resource school curriculum.

Bloom’s-conditioned question generation. A growing line of work conditions LLM question generation on Bloom’s levels. Scaria et al. (2024a) generated questions at each Bloom’s level for the Indian high-school social science curriculum and evaluated them with subject experts using an eight-item rubric, and related work extends this to generation strategies and evaluation (Scaria et al., 2024b). Elkins et al. (2024) showed that teachers can use LLMs together with Bloom’s taxonomy to author quizzes. Using the EduProbe textbook dataset, (Maity et al., 2024) found that the quality of generated questions varies systematically across cognitive levels, with weaker performance at higher-order levels, which is a question-level analogue of the extraction-level suppression we document. These studies operate on English-language curricula and condition each generated question on a single Bloom’s level. We generate questions in Nepali and, motivated by the multi-dimensionality of the objectives, condition generation on both single-label and multi-label taxonomy assignments.

NLP for Nepali and low-resource curricula. Nepali is a low-resource, Devanagari-script lan-

guage for which multilingual models have historically underperformed relative to high-resource languages (Timilsina et al., 2022). Recent work has expanded Nepali resources through benchmark consolidation and generative modelling (Pudasaini et al., 2025), and shared tasks on Devanagari-script languages have established multilingual encoders such as MuRIL (Khanuja et al., 2021) and XLM-RoBERTa (Conneau et al., 2020) as standard backbones (Thapa et al., 2025). More broadly, education is a core application domain of NLP for social good (Karamolegkou et al., 2026), yet work in this space concentrates on high-resource languages and Western curricula, leaving low-resource public school systems underserved. To our knowledge, no prior work studies Bloom’s-conditioned question generation, or the taxonomy assignment it depends on for the Nepali curriculum. We bring the curriculum-aligned generation paradigm to this setting and quantify where its assumed components, single-label objectives and reliable zero-shot taxonomy extraction break down.

Evaluation of generated questions. Automated metrics for question generation largely capture surface linguistic properties and do not assess pedagogical value, so expert evaluation remains standard (Horbach et al., 2020). We adopt this practice, adapting the bilingual pedagogical rubric of Horbach et al. (2020) to accommodate objectives that span multiple cognitive levels, and have practising Nepali-subject teachers rate generated questions.

3 Methodology

Our methodology consists of the following steps. Firstly, we compiled daily learning objectives from the Grade 7 and Grade 10 Nepali subject. The document was in preeti font, so the objectives had to be manually converted to unicode. Then, we had two government secondary Nepali teachers assign revised Bloom’s Revised Taxonomy (Anderson and Krathwohl, 2001) labels to each objectives. They were instructed to assign a single taxonomy per objective first, then were given the option to extend it upto 3. We then zero shot prompted various LLMs to generate the taxonomy and validated whether they agreed with what the teachers assigned. Then we use these labels and objectives to generate grounded questions and had the annotators evaluate their classroom usability with a pedagogical rubric adapted from Horbach et al. (2020), following the adaptation approach of Scaria et al.

(2024a).

3.1 Curriculum Compilation

We compiled daily learning objectives from the Grade 7 and Grade 10 Nepali-subject curriculum published by the Curriculum Development Centre (CDC) of Nepal (Curriculum Development Centre, 2025, 2026). We selected these two grade levels because they represent distinct cognitive demands within the Nepali curriculum: Grade 7 at the lower-secondary level and Grade 10 at the secondary level. The objectives span five chapter types: poem, conversation, story, letter, and biography.

We define a *learning objective* as any statement in the teacher’s guide that specifies the intended cognitive outcome of a lesson, and extracted these objectives manually from the guidelines. In total, we extracted **93** objectives across Grade 7 (**28**) and Grade 10 (**65**).

3.2 Bloom’s Taxonomy Assignment

We assigned taxonomy labels through both human annotators and LLMs, under single-label and multi-label protocols.

Human Annotation. Two annotators, one male and one female participated in taxonomy labelling. Both are practising Government Nepali-subject teachers with 26 years of experience teaching at the lower-secondary and secondary levels. The annotators worked independently with no cross-discussion during labelling. They were initially instructed to assign a single bloom’s level per objective but then were instructed to assign a maximum of three bloom’s level per objectives.

LLM-Based Taxonomy Extraction. We evaluated five LLM-based methods for automated taxonomy assignment, grouped into single-label and multi-label conditions.

Single-label methods. We prompted the model to assign exactly one Bloom’s level per objective using three prompt variants (T1, T2, T3). For T1, we first translated the Nepali objectives to English and then instructed the LLM to extract taxonomy. For T2, we first extracted the verbs after translation and then used those verbs to get the taxonomy and for T3, we directly extracted the taxonomy from the Nepali objectives. All three variants used Gemini Flash 3.5 in a zero shot setting.

Multi-label methods. We designed two multi-label extraction pipelines. In the first (Multi-EN),

objectives written in Devanagari script were translated to English using Gemini Flash 3.5 before being passed to the LLM for taxonomy extraction. In the second (Multi-NP), objectives were processed directly in Nepali with no intermediate translation. For both pipelines the model could assign up to three labels per objective.

Cross-model replication. To assess whether suppression of higher-order taxonomy categories is specific to Gemini Flash, we replicated the Multi-NP pipeline using Gemini Pro, a stronger model from the same family, and two additional model families, Llama 3.3 70B and Qwen 3 32B. All models were evaluated zero-shot with identical prompts. All extraction prompts are provided verbatim in Appendix A (Appendices A.1–A.6).

Classifier Baselines. As a supervised alternative to zero-shot LLM extraction, we trained a linear classification head on top of frozen multilingual encoder representations. We evaluated two encoders: MuRIL (Khanuja et al., 2021), pre-trained on Indian-subcontinent languages including Nepali, and XLM-RoBERTa (Conneau et al., 2020), a cross-lingual encoder trained on 100 languages. Both encoders were kept frozen and only the head parameters were updated. We trained for 10 epochs with the AdamW optimizer at a learning rate of $2e-5$ and a batch size of 8.

3.3 Question Generation

We generated questions for each objective under three conditions. In the *baseline* condition, the model received only the chapter content and generated questions with no taxonomy guidance. In the *single* condition, the model additionally received the learning objective and its assigned Bloom’s taxonomy label as explicit generation constraint, making it focus on only a single objective. For the third, we based the question generation on all the taxonomy that had been assigned to the objective. The baseline and curriculum-guided generation prompts are given in Appendices A.7 and A.8.

3.4 Question Evaluation

The questions were evaluated by the Nepali teachers. We modified three items to reflect the multi-label taxonomy assignment: items that originally assumed a single cognitive demand per question were revised to accommodate objectives spanning multiple Bloom’s levels. The modified rubric is shown in Table 1. Teachers evaluated each question

on all items independently, blind to the generation condition (baseline vs. curriculum-guided), with questions presented in randomised order within each objective. The annotation interface used by the teachers is shown in Figure 3 (Appendix A.9).

Rubric item	Definition
Understandable	Could you understand what the question is asking?
Grammatical	Is the Nepali grammatically correct?
Language Level	Is the language appropriate for this grade?
Clear	Is it clear what the question asks for?
Answerable	Can students reasonably answer this question?
Central	Is this question important for the objective?
Bloom’s Level	What is the Bloom’s skill associated with the question?
Would You Use It?	Would you use this question in class?

Table 1: Hierarchical eight-item rubric used to evaluate questions generated by LLMs.

3.5 Question Generation Application

We created a question generation application based on the methodology described here by collecting objectives throughout grade 7-10 and it is being used in the government school where the annotators teach. We plan to open-source the tool as a result to make it easier for teachers to generate grounded questions and we also plan to create the tool for more classes in the future.

4 Results

4.1 Inter-Annotator Agreement

We calculated the inter-annotator agreement on the assigned single Bloom’s taxonomy. Annotators 1 and 2 agreed on only 44 of 93 objectives (36.7%), yielding a Cohen’s Kappa of $\kappa = 0.13$. On the suggestion of the teachers themselves, we then instructed them to annotate up to 3 taxonomy per objectives. Agreement improved substantially with both annotators assigned at least one common label to 84.6% of objectives, with mean label counts of 2.15 and 2.41 respectively (Table 2). Thus, for single objective, rather than indicating poor annotation quality, low agreement reflects that a single Bloom’s level frequently cannot capture the full

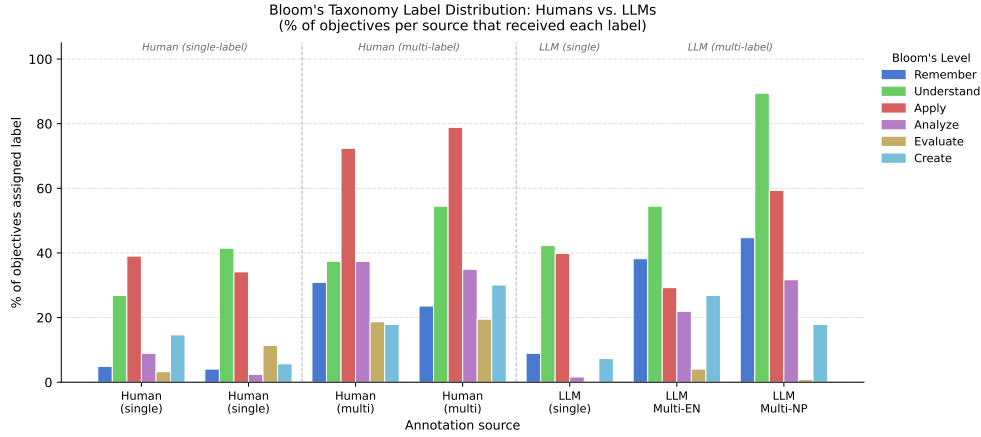


Figure 1: Distribution of Bloom’s taxonomy labels across all annotation sources, expressed as the percentage of objectives assigned to each label.

Metric	Score
Exact set match (%)	18.7
At-least-one match (%)	84.6
Mean Jaccard similarity	0.447
Mean labels (A1_Multi)	2.15
Mean labels (A2_Multi)	2.41

Table 2: Multi-label inter-annotator agreement between Annotators 1 and 2 across 93 objectives.

cognitive demand of objectives in the Nepali curriculum.

The taxonomy label distribution across all the sources is shown in Figure 1.

4.2 LLM Bloom’s Taxonomy Extraction

Table 3 summarises Macro F1 and per-category recall for the multi-label LLM methods, their cross-model replications, and the supervised classifiers; Table 4 reports per-category recall and precision for all five LLM prompting methods against both annotators.

Single-label methods. We compared the performance of the three single-labeled methods against Annotators 1, 2, 1_Multi, and 2_Multi. Among the three single-label methods, LLM-T3 achieves the highest exact match against Annotator 1 at 67.5%, which exceeds the human-human agreement ceiling of 36.7%. Against the richer multi-label human annotations, however, all three methods exhibit high precision (74–83%) but low recall (24–27%) (Table 4), indicating that the LLMs under-generate labels rather than mis-label: when they do assign a category, it tends to be correct, but they systematically miss categories that human annotators identify.

Two systematic failures persist across all three methods regardless of prompting strategy. First, *Evaluate* is never predicted (0% recall) despite Annotators 1 and 2 each assigning it to 23–24 objectives. Second, *Analyze* is severely under-generated, with recall ranging from only 2–13% across methods. Confusion analysis reveals that the most frequent substitution patterns are *Apply*→*Remember* and *Understand*→*Remember*, exposing a consistent bias toward lower-order Bloom’s levels. When *Evaluate* is present, the LLM most commonly replaces it with *Understand* or *Apply*. We attribute this to the surface-level lexical cues available when processing Nepali objectives through zero-shot prompts: lower-order verbs (*describe*, *identify*, *use*) dominate the objectives textually, and without deeper semantic grounding in the curriculum context, the model defaults to the most lexically salient category.

Multi-label methods. Permitting the LLM to assign multiple labels per objective substantially improves alignment with human annotations. Multi-NP outperforms Multi-EN on nearly every metric (Table 4), producing a mean of 2.44 labels per objective against 1.80 for Multi-EN. The performance gap indicates that direct extraction in Nepali preserves cognitive scope that is lost in the translation step. Despite this overall improvement, the suppression of *Evaluate* persists in both multi-label methods, with Multi-EN recovering it only 5 times and Multi-NP only once, suggesting that the failure may be rooted in the model’s conceptual representation of the category rather than in prompt design or language of extraction.

Method	Macro F1	Rec _{Eva}	Rec _{Ana}	Rec _{Rem}	Rec _{Und}	Rec _{App}	Rec _{Cre}
<i>LLM-based methods (zero-shot)</i>							
LLM Multi-EN (Gemini Flash)	0.447	0.079	0.265	0.481	0.576	0.313	0.447
LLM Multi-NP (Gemini Flash)	0.498	0.000	0.397	0.537	0.918	0.609	0.298
LLM Multi-NP (Gemini Pro)	0.459	0.053	0.382	0.500	0.718	0.400	0.319
LLM Multi-NP (Llama)	0.496	0.132	0.338	0.315	0.765	0.652	0.298
LLM Multi-NP (Qwen)	0.443	0.368	0.441	0.019	0.376	0.609	0.340
<i>Classifier over frozen Nepali representations (ours)</i>							
MuRIL	0.550	0.100	0.667	0.556	0.818	0.647	0.500
XLm-RoBERTa	0.612	0.300	0.667	0.667	0.636	0.765	0.875

Table 3: Macro F1 and per-category recall across all methods. Gemini Pro, Llama, and Qwen are evaluated under the same Multi-NP (direct Nepali extraction) protocol as Gemini Flash for comparability. Rec = Recall; Eva = Evaluate; Ana = Analyze; Rem = Remember; Und = Understand; App = Apply; Cre = Create.

Method	Remember		Understand		Apply		Analyze		Evaluate	
	R	P	R	P	R	P	R	P	R	P
<i>Reference: Annotator 1_Multi</i>										
T1	34	39	26	57	25	81	13	46	0	—
T2	18	47	41	46	39	78	9	50	0	—
T3	18	64	54	48	43	78	2	50	0	—
M-EN	42	34	59	40	31	78	22	37	13	60
M-NP	61	42	93	39	62	75	35	41	0	0
<i>Reference: Annotator 2_Multi</i>										
T1	45	39	18	57	24	85	16	54	0	—
T2	21	40	31	51	38	82	12	62	0	—
T3	10	27	40	52	39	78	2	50	0	—
M-EN	62	38	61	61	33	89	30	48	8	40
M-NP	55	29	90	55	61	81	44	49	0	0

Table 4: Per-category Recall (R) and Precision (P) in % for LLM methods against Annotators 1_Multi and 2_Multi. Evaluate recall is 0% across all single-label methods and M-NP; M-EN recovers it partially (13% vs. A1_Multi, 8% vs. A2_Multi). Analyze is similarly suppressed (2–13% recall across single-label methods).

Cross-model replication. To assess whether the suppression of higher-order Bloom’s categories is specific to Gemini Flash, we replicated the Multi-NP extraction pipeline using Gemini Pro, a stronger model from the same family, and two additional model families, Llama 3.3 70B and Qwen 3 32B, whose results are present in Table 3.

Llama reproduces Gemini Flash’s directional failure: *Evaluate* recall reaches only 13.2% and *Analyze* 33.8%, while *Apply* and *Understand* remain the dominant predicted categories. Gemini Pro, despite its larger capacity, shows the same suppression pattern (*Evaluate* recall 5.3%). Qwen exhibits a wildly different failure mode: it partially recovers *Evaluate* (0.368) and *Analyze* (0.441) but nearly eliminates *Remember* (0.019). Across all five LLM configurations, Macro F1 falls within a narrow band of 0.443–0.498, indicating that zero-shot taxonomy extraction approaches a performance ceiling of approximately 0.50 on this task regardless of model family or size. The supervised classifiers break through this ceiling: MuRIL reaches

0.550 and XLm-RoBERTa reaches 0.612 Macro F1, with substantially improved recall for *Analyze* (0.441 → 0.667) and *Create* (0.447 → 0.875), confirming that task-specific supervision recovers the higher-order categories that zero-shot prompting consistently misses. We return to the societal consequences of this suppression for deployed systems in the Ethical Consideration section.

4.3 Question Usability Evaluation

Behavioral finding. Evaluator 2 discontinued assessment of all 20 baseline questions, judging them insufficiently usable to warrant continued evaluation. Baseline results therefore reflect Evaluator 1’s responses only. We treat this discontinuation as a substantive finding rather than data loss: a practicing teacher with 26 years of classroom experience found unguided generated questions categorically unworthy of evaluation, independent of any rubric item.

Teacher evaluation. Figure 2 shows the *Would-Use* distribution across all three conditions. Taxonomy-grounded generation produces a near-complete shift in teacher endorsement relative to the unguided baseline. Evaluator 1 put 12.5% of baseline questions as usable, rising to 81.5% for the single-taxonomy condition and 100% for the multi-taxonomy condition. Evaluator 2 endorsed 93.8% of single-taxonomy questions and 76.5% of multi-taxonomy questions. Averaged across evaluators, endorsement rates are 12.5% (baseline), 87.6% (single), and 88.2% (multi). A Kruskal-Wallis test confirms a highly significant effect of condition ($H = 33.3, p = 5.9 \times 10^{-8}$). Thus teachers highly prefer the taxonomy grounded questions, for both single and multiple taxonomy over questions which were generated without any taxonomy guidance.

Rubric criteria. Table 5 reports evaluator-averaged % Yes responses for each rubric criterion by condition. All generated questions, including baseline, are consistently rated as clear, understandable, grammatically well-formed, answerable, and matching the language level of the students. But a wide gap can be seen on whether the question is central for the objective or not. Baseline questions are rated central by only 50% of responses, whereas single-taxonomy questions reach 96.3% and multi-taxonomy questions reach 100%. Pairwise comparisons confirm that both grounded conditions differ from baseline ($p_{\text{Bonf}} < 0.01$) while single and multi do not differ from each other ($p_{\text{Bonf}} = 1.0$). Taxonomy grounding thus improves usability primarily by anchoring generated questions to the stated objective of the lesson, not by improving their linguistic properties. We note that these linguistic ratings were produced by expert teachers screening every question; we caution against removing that human filter in any adaptation of the pipeline.

Bloom’s taxonomy match. Table 6 reports the match between assigned taxonomy levels and teachers’ perception of the cognitive demand of each generated question. For single-taxonomy questions, overall exact Bloom match (*Yes*) stands at 47.5%, with an additional 49.2% rated *Partially* matching and only 3.4% rated *No*. Match rates vary substantially by assigned level: *Create* achieves the highest exact match (81.8%), while *Analyze* achieves the lowest (22.2%), with 77.8% of *Analyze*-assigned questions rated only *Partially* matching. This repli-

Criterion	Base	Single		Multi		KW p
	E1	E1	E2	E1	E2	
Understandable	100	100	100	100	100	—
Grammatical	88	100	87	100	100	0.22
Lang. Level	88	100	81	100	75	0.95
Clear	88	100	69	100	65	0.93
Answerable	100	100	100	100	100	—
Central	50	100	94	100	100	<0.001

Table 5: Rubric evaluation results by generation condition. % Yes responses are shown per evaluator for conditions where both evaluators participated (single and multi); baseline reflects Evaluator 1 only ($n = 8$) following Evaluator 2’s discontinuation of all baseline evaluations. p -values are from Kruskal–Wallis (KW) tests across conditions; Central is the only criterion with a significant effect of condition.

Level	Single		Multi	
	Yes%	Part.%	W-set%	Multi-lv%
Remember	54.5	45.5	80.0	100.0
Understand	30.0	60.0	68.8	96.9
Apply	44.4	50.0	61.1	100.0
Analyze	22.2	77.8	71.4	92.9
Create	81.8	18.2	50.0	100.0
Evaluate	—*		50.0	—

*Only 1 single-condition question assigned Evaluate.

Table 6: Bloom’s taxonomy match rates by assigned level for single-taxonomy (exact match) and multi-taxonomy (within-set match) generated questions.

cates at the question level the finding that *Analyze* is the most frequently suppressed and misrepresented category in LLM taxonomy extraction: even when the generation prompt specifies *Analyze*, the resulting question does not fully instantiate that cognitive demand as perceived by teachers.

For multi-taxonomy questions, 97.1% were perceived by evaluators as spanning multiple cognitive levels, with a mean of 2.65 labels assigned per question, consistent with the mean label counts of 2.15 and 2.41 observed in Table 2. Overall within-set match reached 70.6%, with *Remember* achieving the highest within-set match (80.0%) and *Create* the lowest (50.0%). For the six Evaluate-assigned multi-taxonomy questions, within-set match stood at 50.0%. The convergence between annotation label counts and perceived label counts suggests that multi-taxonomy generation faithfully reproduces the cognitive complexity of the original objectives.

5 Conclusion

Curriculum guidelines shape classroom practice only when teachers can act on them, and in Nepali

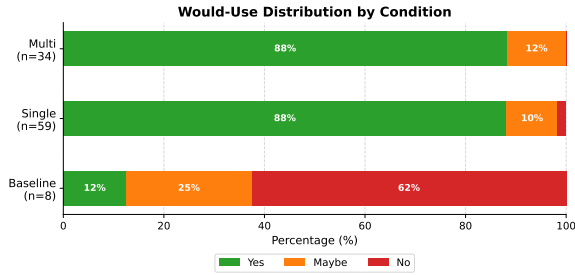


Figure 2: Would-Use distribution by generation condition. Taxonomy grounding, whether single- or multi-label, produces a near-complete shift from unusable to usable questions. Baseline endorsement (*Yes*) stands at 12.5% (Evaluator 1 only; Evaluator 2 discontinued all baseline evaluations), rising to 87.6% for single-taxonomy and 88.2% for multi-taxonomy conditions (Kruskal-Wallis $H = 33.3$, $p = 5.9 \times 10^{-8}$). Single and multi conditions do not differ significantly ($p_{\text{Bonf}} = 1.0$).

classrooms the barrier has been the manual cost of authoring questions aligned with every objective. Our results indicate that this barrier can be removed, but not by prompting an LLM with the curriculum directly. Two design decisions proved essential. First, the taxonomy layer must reflect the structure of the objectives themselves: Nepali learning objectives routinely span several cognitive levels, so single-label annotation collapses to near-chance agreement ($\kappa = 0.13$) while multi-label annotation recovers agreement on at least one level for 84.6% of objectives. Second, the taxonomy labels must come from a reliable source. Zero-shot extraction plateaued at a Macro F1 of 0.443–0.498 across three model families, with each model suppressing a different Bloom’s category, whereas a linear classifier over frozen multilingual encoder representations reached 0.612 and recovered the Analyze and Create categories that prompting missed. When generation was conditioned on these labels, teacher endorsement of the questions rose from 12.5% to 87.6% (single-label) and 88.2% (multi-label), and the improvement was driven almost entirely by whether a question was central to its objective rather than by any linguistic property.

These results also speak to how the social impact of NLP systems should be measured. Technical metrics barely separated the LLM configurations (Macro F1 within 0.443–0.498), yet the difference at the point of use was decisive, and our clearest impact signal was behavioural rather than numeric: an experienced teacher discontinued evaluation of unguided questions entirely. For deployed educa-

tional NLP, practitioner-in-the-loop usability judgments captured what benchmark scores did not, and we suggest they should be treated as a primary impact measure rather than a supplementary one.

The broader lesson for low-resource educational NLP is that the pedagogically critical component of the pipeline, taxonomy assignment, is precisely where zero-shot LLMs fail silently, and that a small amount of task-specific supervision, here 93 annotated objectives, is sufficient to correct it. Question generation itself is not the bottleneck; grounding it is. The resulting application is in use at a government school, and we will release the tool, the annotated objectives, and the generated questions to support extension to further grades and, we hope, to other subjects in the Nepali curriculum and to comparable low-resource school systems.

Limitations

Scale and provenance of annotation. Our gold standard rests on 93 objectives labelled by two practising teachers. While their experience (26 years each) lends the labels authority, two annotators from the same school system cannot capture the full range of defensible taxonomy judgements, and the multi-label agreement we report (exact set match of 18.7%) shows that reasonable experts still differ on the precise label set. The supervised classifiers are trained and evaluated on this same small dataset, so the reported Macro F1 of 0.550–0.612 should be read as evidence that supervision breaks the zero-shot ceiling, not as a stable estimate of deployment accuracy. Evaluate remains the hardest category even for the best classifier (recall 0.3).

Evaluator overlap and baseline coverage. The teachers who annotated the objectives also evaluated the generated questions. Their familiarity with the objectives is what makes the Central criterion meaningful, but it may also bias them toward questions that echo the phrasing of objectives they annotated. In addition, Evaluator 2 discontinued the baseline condition, so baseline rubric results reflect a single evaluator over eight questions. We report the discontinuation as a finding, but it leaves the baseline estimate statistically thin, and the 12.5% baseline endorsement rate should be interpreted with corresponding caution.

Model coverage. Question generation used a single model (Gemini Flash), so the usability results establish that taxonomy grounding helps for this generator rather than for LLM generation in

general. On the extraction side, all LLM methods were run zero-shot; few-shot prompting or instruction fine-tuning might narrow the gap to the supervised classifiers, and the ceiling we report applies to the zero-shot setting specifically.

Outcome measure. Our evaluation captures teacher-perceived usability, not learning outcomes. Whether taxonomy-grounded questions improve student achievement in the Nepali subject requires a classroom study with student-facing measures, which we leave to future work now that the tool is deployed.

Ethical Consideration

What positive impact means in this work. We understand positive impact in three concrete senses. First, workload: authoring diverse, objective-aligned questions is a manual burden that falls hardest on under-staffed government schools, so a tool that generates candidate questions redistributes assessment capacity toward the schools least able to afford it. Second, pedagogical quality: Nepali is the subject with the lowest mean score in the 2023 national assessment ([Education Review Office, 2025](#)), and classroom practice centred on reproducing fixed examination questions is a documented contributor ([Bhattarai and Basnet, 2022](#)); grounding assessment in the curriculum’s own learning objectives is a step away from rote reproduction. Third, alignment with United Nations Sustainable Development Goal 4 (quality education): the intervention targets a public school system in a low-resource language, not a commercial tutoring market. We measure impact at the point of use, through the judgements of the practising teachers who would deploy the questions, rather than through proxy benchmarks alone.

Risk of lowering cognitive demand at scale. Our results identify a risk that is easy to miss when such systems are deployed without scrutiny. Every LLM we tested systematically shifts learning objectives toward lower-order Bloom’s levels, suppressing Evaluate and Analyze (Section 4.2). A question-generation pipeline built naively on zero-shot taxonomy extraction would therefore not be neutral: deployed across a national curriculum, it would quietly reduce the proportion of assessment targeting higher-order thinking, precisely the skills the curriculum reform intends to promote. This failure is silent, because the generated questions remain fluent, grammatical, and answerable; only

their cognitive level is degraded. Our supervised grounding step is the mitigation we adopt, and we regard taxonomy auditing as a necessary safeguard for any comparable deployment, not an optional refinement.

Teacher-in-the-loop design. The tool generates candidate questions for teachers to review, select, and adapt; it does not assess students, assign grades, or interact with students directly. The final pedagogical judgement remains with the teacher, and our evaluation design reflects this: usability was judged by the practitioners who hold that responsibility. We view the tool as reducing the cost of a task teachers already perform rather than replacing teacher expertise, though we acknowledge that sustained use could shift how teachers engage with the curriculum guides, and we plan to observe this in the ongoing deployment.

Data and participants. The learning objectives are drawn from publicly available curriculum documents published by the Curriculum Development Centre of Nepal. No student data was collected at any stage. The two participating teachers gave informed consent for the use of their annotations and evaluations in this study and were compensated for the annotation and evaluation sessions. The deployed application does not log personally identifying information about teachers or students.

Cost, access, and dependence. Our supervised taxonomy classifier uses frozen multilingual encoders with a linear head, so taxonomy assignment can run at negligible cost on modest hardware, which matters for sustainability in a public school system. Question generation, however, currently depends on a commercial LLM API, which introduces recurring cost, network dependence, and exposure to provider changes. We name this honestly as an access risk for the setting we target, and we consider substitution with open-weight models a priority for the open-source release.

Language and cultural fidelity. LLM-generated Nepali may contain subtle grammatical or register errors, and generated questions may reflect the cultural assumptions of predominantly English-centric training data. In our evaluation, teachers rated grammaticality and grade-appropriate language highly across conditions, but this screening was performed by expert teachers; we caution against removing that human filter in any adaptation of the pipeline.

References

- 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.
- Damodar Prasad Bhattarai and Hari Bahadur Basnet. 2022. [Understanding the Nepali classroom practices: A constructivist perspective](#). *Journal of Research and Development*, 5(1):33–40.
- Wen-Chih Chang and Ming-Shun Chung. 2009. [Automatic applying bloom's taxonomy to classify and analysis the cognition level of English question items](#). In *2009 Joint Conferences on Pervasive Computing (JCPC)*, pages 727–733. IEEE.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Curriculum Development Centre. 2025. Shikshak nirdeshika: Kaksha 10 nepali [teacher's guide: Grade 10 nepali]. <https://moecdc.gov.np/content/210/teacher-directory-class-10-nepali-2081-/>. Accessed: June, 2026.
- Curriculum Development Centre. 2026. Shikshak nirdeshika: Kaksha 7 nepali [teacher's guide: Grade 7 nepali]. <https://moecdc.gov.np/content/630/teacher-s-guide-class-7-nepali/>. Accessed: June, 2026.
- Education Review Office. 2025. [National assessment of student achievement 2023: Main report – grade 10 \(mathematics, science, english and nepali\)](#). Technical report, Government of Nepal, Ministry of Education, Science and Technology, Education Review Office, Sanothimi, Bhaktapur, Nepal.
- Sabina Elkins, Ekaterina Kochmar, Jackie C.K. Cheung, and Iulian Serban. 2024. [How teachers can use large language models and bloom's taxonomy to create educational quizzes](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.
- Andrea Horbach, Itziar Aldabe, Marie Bexte, Oier Lopez de Lacalle, and Montse Maritxalar. 2020. [Linguistic appropriateness and pedagogic usefulness of reading comprehension questions](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1753–1762, Marseille, France. European Language Resources Association.
- Antonia Karamolegkou, Angana Borah, Eunjung Cho, Sagnik Ray Choudhury, Martina Galletti, Pranav Gupta, Oana Ignat, Priyanka Kargupta, Neema Kotonya, Hemank Lamba, Sun-Joo Lee, Arushi Mangla, Ishani Mondal, Fatima Zahra Moudakir, Deniz Nazar, Poli Nemkova, Dina Pisarevskaya, Naqee Rizwan, Nazanin Sabri, and 14 others. 2026. [NLP for social good: A survey and outlook of challenges, opportunities and responsible deployment](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5110–5170, Rabat, Morocco. Association for Computational Linguistics.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. [MuRIL: Multilingual representations for Indian languages](#). *arXiv preprint arXiv:2103.10730*.
- Zahra Kolagar, Frank Zalkow, and Alessandra Zarcone. 2025. [Investigating methods for mapping learning objectives to bloom's revised taxonomy in course descriptions for higher education](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 415–445, Vienna, Austria. Association for Computational Linguistics.
- Yuheng Li, Mladen Rakovic, Boon Xin Poh, Dragan Gasevic, and Guanliang Chen. 2022. [Automatic classification of learning objectives based on Bloom's taxonomy](#). In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 530–537, Durham, United Kingdom. International Educational Data Mining Society.
- Subhankar Maity, Aniket Deroy, and Sudeshna Sarkar. 2024. [How effective is gpt-4 turbo in generating school-level questions from textbooks based on bloom's revised taxonomy?](#) *Preprint*, arXiv:2406.15211.
- 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.
- Shushanta Pudasaini, Aman Shakya, Siddhartha Shrestha, Sahil Bhatta, Sunil Thapa, and Sushmita Palikhe. 2025. [Nepaligpt: A generative language model for the nepali language](#). *Preprint*, arXiv:2506.16399.
- Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024a. [How good are Modern LLMs in generating relevant and high-quality questions at different bloom's skill levels for Indian high school social science curriculum?](#) In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 1–10, Mexico City, Mexico. Association for Computational Linguistics.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024b. Automated educational question generation at different bloom’s skill levels using large language models: Strategies and evaluation. In *Artificial Intelligence in Education*, pages 165–179, Cham. Springer Nature Switzerland.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025. [Natural language understanding of Devanagari script languages: Language identification, hate speech and its target detection](#). In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 71–82, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. [NepBERTa: Nepali language model trained in a large corpus](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 273–284, Online only. Association for Computational Linguistics.

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](#). In *Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 209–218, Trento, Italy. Association for Computational Linguistics.

Anwar Ali Yahya, Zakaria Toukal, and Ahmed Osman. 2012. Bloom’s taxonomy–based classification for item bank questions using support vector machines. In *Modern Advances in Intelligent Systems and Tools*, volume 431 of *Studies in Computational Intelligence*, pages 135–140. Springer.

A Appendix

We reproduce verbatim all prompts used in our experiments. Placeholders in curly braces (e.g., {objective_en}) denote runtime substitutions. All prompts were issued zero-shot.

A.1 Translation Prompt

Used to translate objectives from Nepali to English inside the T1, T2, and Multi-EN pipelines.

Translate the following Nepali learning objective into English. Preserve the instructional verbs and the intended cognitive demand exactly; do not simplify, summarise, or add content. Return only the English translation.
Objective: {objective_np}

A.2 T1: Single-Label Extraction from Translated Objective

You are an expert Nepali teacher with many years of experience and you are well versed in bloom’s revised taxonomy. For the objective given as input, you have to assign EXACTLY ONE Bloom’s category that best captures the cognitive demand of the objective.

Constraints:

You have to choose one of the Bloom’s Category between Remember, Understand, Apply, Analyze, Evaluate, Create.

Output:

Respond with only the category name

Input:

The nepali objective for the day, translated into english.

Objective: {objective_en}

A.3 T2: Verb-Mediated Single-Label Extraction

T2 consists of two sequential calls: verb extraction, followed by taxonomy assignment from the extracted verbs.

Step 1: Verb extraction

Below is a learning objective from the Grade {grade} Nepali-subject curriculum, translated into English. List the instructional verb(s) that specify what the student must do. Return only the verbs, comma-separated.

Objective: {objective_en}

Step 2: Taxonomy from verbs

You are an expert Nepali teacher with many years of experience and you are well versed in bloom’s revised taxonomy. For the verbs extracted from an objective given as input, you have to assign EXACTLY ONE Bloom’s category that best captures the cognitive demand of the verbs.

Constraints:

You have to choose one of the Bloom’s Category between Remember, Understand, Apply, Analyze, Evaluate, Create.

Output:

Respond with only the category name

Input:

Verbs from the nepali objective for the day, translated into english.

Objective: {verbs_en}

A.4 T3: Direct Single-Label Extraction from Nepali

You are an expert Nepali teacher with many years of experience and you are well versed in bloom's revised taxonomy. For the objective given as input in Nepali, you have to assign EXACTLY ONE Bloom's category that best captures the cognitive demand of the objective.

Constraints:

You have to choose one of the Bloom's Category between Remember, Understand, Apply, Analyze, Evaluate, Create.

Output:

Respond with only the category name

Input:

The nepali objective for the day, in Nepali.

Objective: {objective_np}

A.5 Multi-EN: Multi-Label Extraction from Translated Objective

You are an expert Nepali teacher with many years of experience and you are well versed in bloom's revised taxonomy. For the objective given as input, you have to assign UP TO THREE Bloom's categories that together capture the full cognitive demand of the objective. Do not assign a category unless the objective genuinely requires it.

Constraints:

You have to choose between one and three of the Bloom's Category between Remember, Understand, Apply, Analyze, Evaluate, Create.

Output:

Respond with only the category names, comma-separated

Input:

The nepali objective for the day, translated into english.

Objective: {objective_en}

A.6 Multi-NP: Multi-Label Extraction Directly in Nepali

The identical prompt was used, unchanged, for all model configurations (Gemini Flash, Gemini Pro, Llama 3.3 70B, and Qwen 3 32B) in the cross-model replication of Section 3.2.

You are an expert Nepali teacher with many years of experience and you are well versed in bloom's revised taxonomy. For the objective given as input in Nepali, you have to assign UP TO THREE Bloom's categories that together capture the full cognitive demand of the objective. Do not assign a category unless the objective genuinely requires it.

Constraints:

You have to choose between one and three of the Bloom's Category between Remember, Understand,

Apply, Analyze, Evaluate, Create.

Output:

Respond with only the category names, comma-separated

Input:

The nepali objective for the day, in Nepali.

Objective: {objective_np}

A.7 Baseline Question Generation

You are an expert Nepali teacher with many years of experience. For the chapter given as input, you have to generate {N} questions in Nepali that a teacher could use in class for this chapter.

Constraints:

The questions must be answerable from the chapter content.

Output:

Respond with only the questions, numbered.

Input:

The chapter from the Grade {grade} Nepali textbook.

Chapter ({chapter_type}): {chapter_content}

A.8 Curriculum-Guided Question Generation

Used for both the single-taxonomy and multi-taxonomy conditions; in the single condition {taxonomy_labels} contains one level, and in the multi condition it contains the full assigned label set.

You are an expert Nepali teacher with many years of experience and you are well versed in bloom's revised taxonomy. For the chapter, learning objective, and assigned Bloom's level(s) given as input, you have to generate {N} questions in Nepali.

Constraints:

The questions must be answerable from the chapter content.

The questions must directly serve the stated learning objective.

The questions must require the cognitive process(es) named in the assigned Bloom's level(s).

Output:

Respond with only the questions, numbered.

Input:

The chapter from the Grade {grade} Nepali textbook, the learning objective from the Curriculum Development Centre teacher's guide, and its assigned Bloom's taxonomy level(s).

Chapter ({chapter_type}): {chapter_content}

Learning objective: {objective_np}

Assigned Bloom's level(s): {taxonomy_labels}

A.9 Annotation Tool

The interface given to teachers for question evaluation is shown in Figure 3.

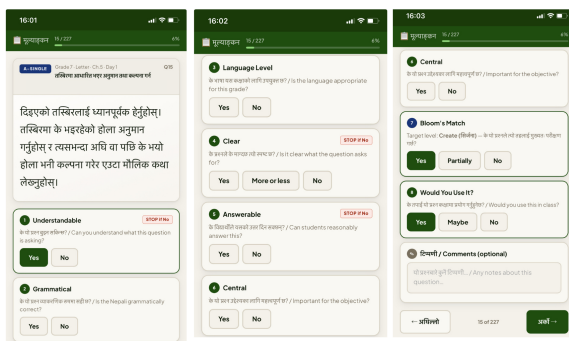


Figure 3: Annotation interface given to teachers for question evaluation.