Bridging the Multilingual Gap in Educational Question Generation via Knowledge Distillation

Byambaa Bayarmandakh [1] **Sahan Bulathwela** [1 2]

Abstract

Question Generation (QG) is central to Intelligent Tutoring Systems, but routine Large Language Model (LLM) inference is difficult to deploy in resource-constrained educational settings, including much of the Global South. We study whether sequence-level knowledge distillation can transfer multilingual, topic-controlled QG from a teacher LLM into a deployable Small Language Model (SLM) without commissioning per-language datasets[1]. Using XQuAD contexts in 11 languages, Google Gemini 2.5 Flash generates synthetic training pairs and an mT5-small student is fine-tuned on the result. Against zero-shot LLMs and a bespoke augmented mT5 baseline, the distilled student retains 86–87% of the teacher's lexical and WikiSemRel topic-control performance while using a far smaller model. The results suggest that distillation is a practical pathway for multilingual educational NLP under limited data and compute.

1. Introduction

Global South educational settings foreground machine learning under resource constraints: limited data, limited compute, and deployment settings where large proprietary systems are hard to run continuously. Educational Question Generation (QG) is a concrete instance of the challenge. Intelligent Tutoring Systems can use QG to scale practice and feedback, but multilingual educational data is scarce, and frontier LLM inference is costly for many institutions in the Global South (Lhaksmana et al., 2024).

The technical question is therefore not only whether LLMs can generate good questions, but whether their capabilities can be transferred into smaller models that local institutions can run more cheaply and repeatedly. English educational QG has seen strong progress with small T5-style models (Bulathwela et al., 2023; Fawzi et al., 2024; Li et al., 2025); however, those systems rely on resources such as SQuAD and KhanQ that are much richer than the datasets available for many languages. A central mismatch follows: the case for efficient SLM deployment is strongest in precisely the settings where supervised multilingual training data is weakest.

We study sequence-level knowledge distillation (Hinton et al., 2015; Sanh et al., 2019) as a practical way to bridge the gap. A multilingual teacher LLM generates topic-controlled questions (Li et al., 2025) from XQuAD contexts (Artetxe et al., 2020) in 11 languages, and an mT5-small (Xue et al., 2020) student is fine-tuned on the synthetic pairs. The result is a low-resource recipe for Global South educational settings, combining synthetic data generation, model compression, multilingual NLP, and educational impact under deployment constraints.

Recent QG surveys flag two recurring gaps in the literature. Lexical-overlap metrics remain dominant despite well-documented limitations (Guo et al., 2024; Dong et al., 2025; Mohammadshahi et al., 2023), and multilingual educational QG is under-treated even though the underlying LLMs are themselves reliably multilingual (Dong et al., 2025; Qin et al., 2025). Augmentation over small bespoke corpora is also susceptible to n-gram overfitting that lexical metrics reward and semantic metrics do not (Nguyen et al., 2024). We organise our study around four research questions targeting these gaps: **(RQ1)** can knowledge distillation scale SLM performance for QG in multilingual settings; **(RQ2)** how large is the teacher→student degradation when distilling a multilingual LLM into mT5-small; **(RQ3)** does distillation outperform training on bespoke augmented multilingual datasets under semantic (rather than purely lexical) evaluation; **(RQ4)** do reference-free metrics (WikiSemRel, RQUGE-LLM) yield different model rankings than lexical metrics? Extended background, related work, and methodology are deferred to Appendix A.

[1]Department of Computer Science, University College London, London, United Kingdom [2]Centre for Artificial Intelligence, University College London, London, United Kingdom. Correspondence to: Byambaa Bayarmandakh <byambaa.bayarmandakh.25@ucl.ac.uk>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

[1]The dataset, code, and full experiment records will be available on GitHub upon archival.

2. Background and Motivation

Educational QG with SLMs. Prior work shows that compact models can be competitive for educational QG in English. Bulathwela et al. (Bulathwela et al., 2023) introduce EDUQG, a T5-small model adapted to scientific educational text. Fawzi et al. (Fawzi et al., 2024) show that a lightweight SLM pipeline can match or exceed larger baselines under several automatic and human-evaluation measures. Li et al. (Li et al., 2025) add topic control through SQuAD/MixSQuAD2X-style augmentation and WikiSemRel evaluation. These results motivate SLMs for low-cost deployment albeit limited to English-only.

Multilingual data scarcity. Multilingual QG inherits the usual low-resource NLP bottleneck: high-quality parallel or language-specific educational question datasets are expensive to commission, especially for communities already underserved by ML infrastructure. XQuAD (Artetxe et al., 2020) offers a useful multilingual benchmark, but its scale remains small relative to English corpora. Teacher-generated synthetic data is therefore attractive: it converts an expensive, centralized model into a one-time data generator for a cheaper student.

Evaluation beyond overlap. Lexical metrics remain useful for tracking reference similarity, but they can reward surface copying and penalize valid paraphrases (Mohammadshahi et al., 2023; Nguyen et al., 2024). We therefore pair lexical metrics with WikiSemRel (Li et al., 2025), which measures entity-level topic relevance, and an RQUGE-style answerability score (Mohammadshahi et al., 2023). Multilingual QG requires distinction because a model may be answerable but poorly topic-controlled, or topic-relevant but lexically distant from a single reference.

3. Method

Task. Following Li et al. (Li et al., 2025), we define topic-controlled QG as generating a question Q given a context paragraph C and target topic T. All students use the same input format, `<topic> {T} <context> {C}`, and are trained to emit Q. The student backbone is mT5-small (Xue et al., 2020) (≈300M parameters), which keeps the deployment target far smaller than the teacher. Training hyperparameters are listed in Appendix I.

Distillation pipeline. The teacher model is Google Gemini 2.5 Flash (Comanici et al., 2025). For each XQuAD train/validation paragraph, the teacher first extracts pedagogically relevant topics and then generates one question per topic in the paragraph's source language (full prompts in Appendix J). We train three variants: MT5 GEMINI DISTIL 1X, 2X, and 3X, using one, two, and three synthetic questions per paragraph respectively. XQuAD test contexts are excluded from both teacher prompting and student training.

Table 1. XQuAD main results. Lexical scores are macro-averaged over 11 languages. WikiSemRel is measured on the English slice; RQUGE-LLM is the 11-language answerability score.

Model	F1	R-L	WSR $\hat{q}_t$	WSR Δ	RQUGE
mT5 MixXQuAD2X	**0.435**	**0.281**	0.269	0.221	**2.496**
Aya Expanse 8B	0.303	0.180	–	–	2.386
Qwen 2.5 7B	0.283	0.161	–	–	2.168
Gemini 2.5 Flash teacher	0.302	0.168	**0.675**	**0.605**	2.111
mT5 Gemini Distil 3x	0.261	0.143	0.584	0.518	1.442
mT5 Gemini Distil 2x	0.259	0.147	0.485	0.358	1.420
mT5 Gemini Distil 1x	0.266	0.145	0.334	0.239	1.327
mT5 zero-shot	0.024	0.022	–	–	1.022

Data and baselines. XQuAD (Artetxe et al., 2020) is the main multilingual benchmark. We evaluate 11 languages: Arabic, German, Greek, English, Spanish, Hindi, Romanian, Russian, Thai, Turkish, and Vietnamese. Chinese is excluded because the teacher failed to reliably generate usable Chinese distilled questions in our pipeline. For comparison, we train MT5 MIXXQUAD2X, a multilingual adaptation of Li et al.'s SQuAD/MixSQuAD2X setup (Li et al., 2025), and evaluate zero-shot Aya Expanse 8B (Dang et al., 2024) and Qwen 2.5 7B (Qwen et al., 2025). KhanQ (Gong et al., 2022) is used as an English educational generalisation benchmark.

Metrics. We report F1, METEOR, and ROUGE-L for lexical overlap. For WikiSemRel, WAT-linked entity sets E_q and E_r define

$$\hat{q}_t = \begin{cases} 1, & \text{if } E_q \cap E_r \neq \varnothing, \\ \mathrm{Jaccard}(E_q, E_r), & \text{otherwise,} \end{cases} \tag{1}$$

with distractor contrast $\Delta = \hat{q}_t - \hat{q}_{t'}$ (Li et al., 2025). Because WAT (Piccinno & Ferragina, 2014) has limited non-English coverage, WikiSemRel is measured on English slices. For answerability, RQUGE-LLM follows the RQUGE architecture (Mohammadshahi et al., 2023): $a_c = \mathrm{QA}(q_c, D)$ and $\kappa = S(q_c, a_c, a_r, D)$, with XLM-R as QA reader and Gemini 2.5 Flash as judge.

4. Results

Lexical performance. MIXXQUAD2X tops the lexical ranking (Table 1; F1 0.435, ROUGE-L 0.281), well above the teacher (F1 0.302) and the zero-shot Aya (0.303) and Qwen (0.283). The pattern is consistent with lexical overfitting: MIXXQUAD2X learns the n-gram distribution of the augmented XQuAD reference set rather than the underlying topic→question mapping — the same failure mode Nguyen et al. (Nguyen et al., 2024) report for reference-based QG metrics. The distilled students sit between the teacher and the zero-shot mT5 floor (0.024 F1), showing the synthetic

supervision is informative. Per-language scores appear in Appendix G.

Topic relevance. The lexical ranking reverses under WikiSemRel. The teacher is strongest on prescribed-topic relevance ($\hat{q}_t = 0.675$, $\Delta = 0.605$), and the distilled 3x student reaches $\hat{q}_t = 0.584$, $\Delta = 0.518$ — essentially matching Claude Sonnet 4.6 on KhanQ (0.585/0.502; Table 2). MIXXQUAD2X, despite topping the lexical ranking, falls to $\hat{q}_t = 0.269$, below every LLM we tested. The 1x student also drops to $\hat{q}_t = 0.334$, $\Delta = 0.239$, indicating that the extra teacher generations in 3x add useful semantic diversity rather than near-duplicates: the bespoke baseline reproduces surface forms while distillation transfers the teacher's topic-selection behaviour.

Answerability. RQUGE-LLM favours MIXXQUAD2X (2.496), then Aya (2.386) and Qwen (2.168), with the teacher at 2.111 and the distilled student lower still (3x: 1.442; 1x: 1.327; per-language scores in Appendix H). The ranking is driven by the metric's reliance on gold answer spans: the SQuAD-style XLM-R reader is biased toward retrieving the reference span, and MIXXQUAD2X's reference-imitating outputs align well with that retrieval target. The distilled student shows the opposite pattern — lower RQUGE-LLM but a 2.2× larger WSR — indicating questions that address the prescribed topic from formulations the reader retrieves less reliably. We therefore treat RQUGE-LLM as an answerability signal alongside, not in place of, WSR.

Metrics measure different qualities (RQ4). Across the KhanQ model suite ($n = 13$), WSR and RQUGE-LLM are positively but only moderately correlated (Pearson $r = 0.60$, $p = 0.029$; Spearman $\rho = 0.64$, $p = 0.019$); $\rho^2 \approx 41\%$ of model-rank variance is shared, leaving enough unshared variance to support within-band rank inversions such as MIXXQUAD2X's RQUGE-LLM lead alongside its low WSR. WSR tracks recall-weighted lexical metrics tightly (F1 ρ=0.89, ROUGE-L ρ=0.87, both $p<0.001$), whereas RQUGE-LLM has only one significant Spearman correlation with any lexical metric (F1 ρ=0.62, p=0.025). A within-model prompt-optimisation study reinforces the separation: holding the LLM fixed, an optimised prompt raises Claude HAIKU's WSR by $+0.050$ (McNemar $p<0.001$) while reducing its RQUGE-LLM by -0.081 (paired t, p=0.024). A perturbation that improves topic grounding while degrading answerability is inconsistent with the two metrics measuring the same underlying signal (full analysis in Appendices L and K).

English generalisation. Table 2 shows that the multilingual distilled student also transfers to KhanQ, an English educational benchmark. It does not surpass the English-specialised Li et al. (Li et al., 2025) reference model, but its WikiSemRel score is close to Claude Sonnet 4.6 and well above Qwen 2.5 7B. The gap to stronger LLM baselines is small under WikiSemRel, so the multilingual training does not visibly trade off English topic control.

Table 2. English KhanQ topic relevance. The distilled multilingual student remains competitive with stronger English LLMs under WikiSemRel.

Model	F1	R-L	WSR $\hat{q}_t$	WSR Δ
Li et al. (Li et al., 2025) T5 topicQG2X	0.321	0.216	0.735	0.680
Gemini 2.5 Flash	0.275	0.178	**0.675**	**0.605**
mT5 Gemini Distil 3x	0.262	0.156	0.584	0.518
mT5 Gemini Distil 2x	0.273	0.189	0.485	0.358
Claude Sonnet 4.6	0.272	0.172	0.585	0.502
Qwen 2.5 7B	0.251	0.163	0.405	0.243

Distillation vs. bespoke training (RQ1–RQ3). Across the 11 evaluation languages, the 3x student retains 86% of the teacher's F1 (0.261 vs. 0.302), 85% of its ROUGE-L (0.143 vs. 0.168), and 87% of its WSR $\hat{q}_t$ (0.584 vs. 0.675) at $\sim$23× fewer active parameters — addressing RQ1 and bounding the teacher→student degradation for RQ2. Relative to the bespoke MIXXQUAD2X baseline, distillation trades lexical overlap for a 2.2× larger WSR ($\hat{q}_t$: 0.584 vs. 0.269; Δ: 0.518 vs. 0.221), addressing RQ3: augmentation over XQuAD's ≈1,190 paragraphs per language approximates n-gram averages of the references, while distillation, generated fresh per context at 3× scale, transfers more of the teacher's topic-selection behaviour. WSR rises monotonically with the per-paragraph generation count k (KhanQ $\hat{q}_t$: 1x 0.334, 2x 0.485, 3x 0.584; Δ: 0.239, 0.358, 0.518), supporting the diversity attribution; the 86–87% retention figures above are k=3-specific (k=1 retains ≈50% of the teacher's WSR). Per-language scores show the gap concentrated in morphologically rich or non-space-tokenised languages (Thai, Arabic, Hindi) compared to Indo-European ones (Appendix G).

5. Resource-Constrained Deployment in the Global South

L et al. (Lhaksmana et al., 2024) identify four barriers to adopting LLMs in Global South universities: ethical misuse, data privacy and security, reliability and accuracy, and cost. These concerns are directly relevant to educational QG, where institutions need generated practice questions to be affordable, auditable, accurate enough for instruction, and compatible with local rules for student and classroom data.

Local distillation turns these constraints into design choices. The teacher LLM is used once to construct synthetic supervision; the resulting student is mT5-small (∼300M parameters, ∼23× smaller than the 7–8B zero-shot baselines we tested) and was trained end-to-end on a single 8 GB laptop

GPU (Appendix I). After that one-shot teacher cost, routine classroom generation can stay within institutional infrastructure rather than being sent to an external proprietary service, in line with the SLM deployment argument of Fawzi et al. (Fawzi et al., 2024).

The governance advantage is especially important for reliability and quality assurance. The distilled model is weaker than the teacher, yet it retains most of the teacher's lexical and WikiSemRel topic-control performance and outperforms the bespoke augmented baseline on topic relevance. On KhanQ, the student essentially matches Claude Sonnet 4.6 on prescribed-topic relevance (0.584 vs. 0.585) and exceeds it on WSR Δ topic separation (0.518 vs. 0.502), showing that a trained local SLM can outperform proprietary LLM access on targeted educational criteria. Because the model is local and task-specific, educators can inspect generated questions, remove inappropriate outputs, record known failures, and continue fine-tuning on smaller curated datasets. The resulting human–AI workflow can scale multilingual question banks while producing further curated data for model improvement.

The comparison with zero-shot LLMs clarifies the deployment trade-off. Aya Expanse 8B and Qwen 2.5 7B provide strong multilingual baselines without task-specific training, but repeated serving remains substantially more expensive than serving mT5-small and offers less local control over the deployed system. The practical claim is therefore optimistic but bounded: distillation can make multilingual educational QG more governable by shifting repeated generation from proprietary model access to a smaller, inspectable model that local institutions can adapt to their curricula, policies, infrastructure, and future intelligent tutoring services.

6. Discussion

Knowledge distillation addresses two Global South bottlenecks at once: it shifts repeated inference from a frontier LLM to a $\sim$23$\times$ smaller mT5-small student that retains 87% of the teacher's WSR topic control, and it replaces bespoke language-specific dataset construction with a one-shot multilingual teacher pass. The approach is not a substitute for local dataset development but provides a practical bootstrap path for under-resourced educational NLP.

The broader design pattern is reusable: use an expensive model to create supervision under controlled prompts, then train a smaller open model for repeated use. The pattern is most attractive when the task can be specified clearly, outputs can be filtered, and the student is evaluated on dimensions that matter for deployment. Topic-controlled QG satisfies these conditions better than unconstrained educational text generation.

The results also caution against single-metric evaluation in multilingual QG. The model with the strongest F1 is not the model with the strongest WSR topic control; WSR and RQUGE-LLM share only $\rho^2 \approx 41\%$ of their model-rank variance, and a within-model prompt change can move the two in opposite directions (Appendix K). The currently best topic-relevance and answerability metrics are also originally English-only (Li et al., 2025; Mohammadshahi et al., 2023), so a fully multilingual semantic-topic and answerability suite remains an open problem.

Limitations. The pipeline excludes Chinese because Gemini 2.5 Flash failed to reliably generate distilled Chinese questions in our setup, plausibly due to provider-side language constraints; we therefore report on the remaining 11 XQuAD languages. WSR depends on WAT (Piccinno & Ferragina, 2014) entity linking with limited non-English coverage, so $\hat{q}_t$ and Δ are reported only on English slices. RQUGE-LLM uses Gemini 2.5 Flash as judge and is biased toward gold-span retrieval, which favours bespoke models trained on XQuAD references; we therefore report it alongside, not instead of, WSR. Finally, our pipeline uses sequence-level data distillation with no access to teacher logits or intermediate states; logit-matching, layer alignment (Jiao et al., 2020), or temporally adaptive interpolation (Shing et al., 2025) could narrow the teacher$\rightarrow$student gap further, especially for low-resource languages where synthetic data is sparser.

7. Conclusion

We show that sequence-level distillation from a multilingual teacher can transfer topic-controlled educational QG into mT5-small with bounded teacher$\rightarrow$student degradation (RQ1–RQ2: 86% F1, 87% WSR retention at $\sim$23$\times$ fewer parameters) and stronger semantic topic control than a bespoke augmented baseline (RQ3: 2.2$\times$ larger WSR). The cross-metric analysis (RQ4) shows that lexical overlap, answerability, and topic relevance capture distinct dimensions of question quality — their cross-model rank correlation is only $\rho^2 \approx 41\%$ — so single-metric reporting is insufficient for multilingual educational QG. For Global South settings, the findings support a practical recipe: use a powerful teacher once to create multilingual supervision, then deploy a smaller, locally inspectable student for lower-cost educational generation, with WSR and RQUGE-LLM reported jointly to surface the topic-control vs. answerability trade-off.

Acknowledgments We thank the reviewers for their reviews and Max Norris for valuable feedback on the research questions. This work is partially funded by the European Commission's projects "Teacher-AI Complementarity (TaiCo)" (Project ID: 101177268) and "Humane AI" (Grant No. 820437) and "X5GON" (Grant No. 761758).

References

Artetxe, M., Ruder, S., and Yogatama, D. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4623–4637, 2020. doi: 10.18653/v1/2020.acl-main.421.

Bulathwela, S., Muse, H., and Yilmaz, E. Scalable educational question generation with pre-trained language models. In *Proceedings of the 24th International Conference on Artificial Intelligence in Education (AIED)*, 2023. URL https://arxiv.org/abs/2305.07871.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., Kublik, S., Amer, M., Aryabumi, V., Campos, J. A., Tan, Y.-C., Kocmi, T., Strub, F., and others. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL https://arxiv.org/abs/2412.04261.

Dong, X., Zhang, X., Li, Z., Hou, Q., Xue, J., and Li, X. A literature review of research on question generation in education. *PeerJ Computer Science*, 2025. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12453861/.

Fawzi, F., Balan, S., Cukurova, M., Yilmaz, E., and Bulathwela, S. Towards human-like educational question generation with small language models. In *Proceedings of the 25th International Conference on Artificial Intelligence in Education (AIED)*, 2024. URL https://discovery.ucl.ac.uk/id/eprint/10196728/.

Gong, H., Pan, L., and Hu, H. KHANQ: A dataset for generating deep questions in education. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H. (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5925–5938, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.518/.

Guo, S., Liao, L., Li, C., and Chua, T.-S. A survey on neural question generation: Methods, applications, and prospects. *arXiv preprint arXiv:2402.18267*, 2024. URL https://arxiv.org/abs/2402.18267.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics (EMNLP)*, pp. 4163–4174, 2020. doi: 10.18653/v1/2020.findings-emnlp.372.

Lhaksmana, K. M., Ishida, T., Ihsan, A. F., and Rudawan, R. A. Advancing global south university education with large language models, 2024. URL https://arxiv.org/abs/2410.07139.

Li, Z., Cukurova, M., and Bulathwela, S. A novel approach to scalable and automatic topic-controlled question generation in education. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK'25)*, 2025. doi: 10.1145/3706468.3706487. URL https://arxiv.org/abs/2501.05220.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. S2ORC: The semantic scholar open research corpus. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://aclanthology.org/2020.acl-main.447/.

Mohammadshahi, A., Scialom, T., Yazdani, M., Brunet, M., Henderson, J., Scialom, G., and Foster, J. RQUGE: Reference-free metric for evaluating question generation by answering the question. *Findings of ACL*, 2023. URL https://aclanthology.org/2023.findings-acl.428/.

Nguyen, B., Yu, M., Huang, Y., and Jiang, M. Reference-based metrics disprove themselves in question generation. In *Findings of the Association for Computational Linguistics (EMNLP)*, 2024. URL https://arxiv.org/abs/2403.12242.

Piccinno, F. and Ferragina, P. From tagme to wat: a new entity annotator. In *Proceedings of the First International Workshop on Entity Recognition & Disambiguation*, ERD '14, pp. 55–62, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450330237. doi: 10.1145/2633211.2634350. URL https://doi.org/10.1145/2633211.2634350.

Qin, L., Chen, Q., Zhou, Y., Chen, Z., Li, Y., Liao, L., Li, M., Che, W., and Yu, P. S. A survey of multilingual large language models. *Patterns*, 6(1):101118, 2025. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2024.101118. URL https://www.sciencedirect.com/science/article/pii/S2666389924002903.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Shing, M., Misaki, K., Bao, H., Yokoi, S., and Akiba, T. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models, 2025. URL https://arxiv.org/abs/2501.16937.

Welbl, J., Liu, N. F., and Gardner, M. Crowdsourcing multiple choice science questions. *CoRR*, abs/1707.06209, 2017. URL http://arxiv.org/abs/1707.06209.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934, 2020. URL https://arxiv.org/abs/2010.11934.

A. Extended Background and Research Questions

Intelligent Tutoring Systems (ITS) rely on automated Question Generation (QG) to scale educational access. Recent surveys document a rich catalogue of neural QG architectures, control signals, and pedagogical applications. Guo et al. (Guo et al., 2024) taxonomise neural QG across structured, unstructured, and hybrid input modalities; Dong et al. (Dong et al., 2025) focus on educational QG specifically, tracing diversification across cognitive level, subject, difficulty, and target topic. Two gaps recur across the literature. First, lexical-overlap metrics (BLEU, ROUGE) remain the de facto standard despite well-documented limitations (Mohammadshahi et al., 2023); both surveys note that answerability and diversity are under-developed, and semantic alternatives such as BERTScore are only recently gaining traction. Second, multilingualism remains under-treated (Dong et al., 2025); the authors state that their review "primarily focuses on studies related to English QG", reflecting the field at large, even though the underlying LLMs have themselves become reliably multilingual (Qin et al., 2025).

These gaps are most consequential in deployment-constrained settings. Routine Large Language Model (LLM) inference is often impractical in much of the Global South (Lhaksmana et al., 2024), where the demand for automated instruction is also high. The prevailing response has been to train Small Language Models (SLMs) that remain competitive at lower cost (Fawzi et al., 2024). Bulathwela et al. (Bulathwela et al., 2023) introduce EDUQG, a T5-small (60M parameters) further pre-trained on scientific abstracts (S2ORC) and fine-tuned on SciQ (Welbl et al., 2017), establishing the SLM backbone for educational QG. Fawzi et al. (Fawzi et al., 2024) extend this with a lightweight grammar corrector and a human-evaluation study, showing the resulting SLM matches a $4\times$ larger LLM baseline and is preferred to (or rated equal to) human-written questions by 60% of raters. Li et al. (Li et al., 2025) then add topic as an explicit control variable, training T5 with dataset-mixing augmentation (MIXSQUAD / MIXSQUAD2X) and grounding evaluation in the WikiSemRel semantic-topic metric. All three contributions, however, are restricted to English: foundational corpora such as SQuAD (Rajpurkar et al., 2016) and KhanQ (Gong et al., 2022) are substantially larger than their multilingual counterparts, e.g. XQuAD (Artetxe et al., 2020), and dataset construction for low-resource educational settings is itself an open problem identified by Dong et al. (Dong et al., 2025). The economic case for SLMs thus arises in the linguistic settings where the supporting data is least available.

Evaluation poses a related challenge. Traditional lexical metrics do not adequately capture semantic relevance or answerability (Mohammadshahi et al., 2023). Semantic metrics such as WikiSemRel (Li et al., 2025) align more closely with human judgement, but their reliance on Wikipedia-backed entity linkers limits coverage in low-resource multilingual settings. A reliable evaluation must therefore combine lexical and semantic dimensions.

Knowledge distillation offers a practical alternative to per-language dataset construction: a multilingual teacher LLM can generate training data in any target language, avoiding the cost of commissioning language-specific corpora. On a narrowly defined task such as topic-controlled QG, distillation may also yield more diverse training signal than augmentation. Augmentation over small bespoke corpora is susceptible to n-gram overfitting that lexical metrics reward and semantic metrics do not, a pattern Nguyen et al. (Nguyen et al., 2024) report on standard QG benchmarks. Distillation, in contrast, draws its diversity from the teacher's output distribution rather than from a fixed pool of references. The four research questions stated in the main introduction (RQ1–RQ4) formalise this stance.

B. Extended Related Work

Surveys of QG and the multilingual gap. Recent surveys frame the QG landscape. Guo et al. (Guo et al., 2024) taxonomise neural QG into structured, unstructured, and hybrid input modalities, surveying RNN, Transformer, graph- and pretrained language model based architectures and explicitly flagging that n-gram metrics "can potentially penalize well-formed questions that diverge in lexical similarity from the ground-truth" and that diversity metrics like Distinct-n are "overly simplistic". Dong et al. (Dong et al., 2025) focus specifically on educational QG, tracing the diversification of generation across cognitive level, subject, difficulty, and topic, and identifying the construction of metadata-rich educational datasets as a critical bottleneck. Notably, Dong et al. explicitly scope their review to English ("the majority of the literature retrieved from citation databases centers on English"), and Guo et al. likewise treat multilingual QG only in passing as one application of cross-lingual transfer. Multilingual educational QG, in particular, sits in the intersection that neither survey covers, which is the focus of this work.

Small Language Models for educational QG. LLMs achieve strong performance on generation tasks, but their latency and cost are often prohibitive for educational deployments. Three closely related contributions establish the SLM line we extend. Bulathwela et al. (Bulathwela et al., 2023) introduce EDUQG, a T5-small (60M parameters) further pre-trained

on the S2ORC (Lo et al., 2020) scientific-abstracts corpus and fine-tuned on SciQ, and establish the SLM-for-EdQG baseline. Fawzi et al. (Fawzi et al., 2024) extend EDUQG with a RoBERTa-based GECToR grammar corrector, showing the resulting stack matches the 4× larger T5-base "Leaf" baseline on SciQ and outperforms ChatGPT on several lexical and semantic measures, with human raters preferring the SLM's outputs to (or rating them equal to) human-written questions in 60% of pairwise comparisons. Li et al. (Li et al., 2025) then add topic as an explicit control variable: they train T5 with a dataset-mixing augmentation strategy (MIXSQUAD / MIXSQUAD2X) and ground evaluation in the WikiSemRel semantic-topic metric. All three contributions are English-only. We examine whether the SLM-competitiveness claim extends to multilingual settings without relying on language-specific bespoke datasets.

Evaluation Metrics. Lexical overlap metrics (BLEU, ROUGE) are known to be unreliable for QG. Mohammadshahi et al. (Mohammadshahi et al., 2023) document these limitations and introduce RQUGE, a reference-free metric that scores the answerability of a generated question. Nguyen et al. (Nguyen et al., 2024) extend the argument: on standard QG benchmarks, reference-based metrics fail to rank human-written questions as competitive. WikiSemRel (Li et al., 2025) provides a complementary semantic signal by measuring how close the Wikipedia entities evoked by the generated question are to those of the target topic. We combine all three families (lexical, answerability-based, and topic-relevance) to evaluate cross-lingual generation along complementary dimensions.

Knowledge Distillation for NLP. Knowledge distillation (Hinton et al., 2015) has been widely used to compress LLM capabilities into smaller models (Sanh et al., 2019; Jiao et al., 2020). Our setting differs in two respects. First, we use the teacher as a data generator for a structured, topic-controlled task, a form of sequence-level distillation that requires neither teacher logits nor a shared tokeniser. Second, we deploy specifically as a substitute for bespoke multilingual data collection, framing distillation as the answer to data scarcity rather than to model size alone.

C. Extended Methodology

The pipeline is designed to isolate the effect of knowledge distillation against matched, augmented baselines.

C.1. Task Formulation

We define topic-controlled QG as the probability of generating a question Q given a context paragraph C and a target topic T:

$$P(Q \mid C, T) = \prod_{i=1}^{|Q|} P(q_i \mid C, T, q_{<i}), \tag{2}$$

where q_i is the i-th token of the generated question. All student models share the input format `<topic>` {T} `<context>` {C} and are trained to emit Q.

C.2. Baselines and Augmented Control

As baselines we evaluate a suite of frontier and open-weight LLMs zero-shot on XQuAD and KhanQ. To extend the SLM approach of (Li et al., 2025) to a multilingual setting, we train an mT5-small(Xue et al., 2020) model on XQuAD and apply Li et al.'s data-augmentation strategy, producing MIXXQUAD2X. It serves as the primary control for SLM training without distillation.

C.3. Knowledge Distillation Pipeline

To address the limited availability of multilingual training data, we distill from the teacher on XQuAD train+val contexts only (test is never shown to the teacher). For each context, the teacher first identifies k pedagogically relevant topics, then generates one question per topic in the paragraph's source language, yielding a $k\times$ synthetic dataset per language. We train three student variants: MT5 (GEMINI DISTIL 1X), (2X), and (3X), corresponding to $k \in \{1, 2, 3\}$ questions per paragraph. The exact prompts and decoding parameters appear in Appendix J. Because XQuAD-test is unused during distillation and training, the per-language evaluation (Table 3, Appendix G) constitutes a held-out generalisation test for both teacher and student.

D. Extended Experimental Setup

Datasets. XQuAD (Artetxe et al., 2020) is our primary multilingual benchmark and distillation source. We evaluate on 11 of the 12 XQuAD languages (Arabic, German, Greek, English, Spanish, Hindi, Romanian, Russian, Thai, Turkish, Vietnamese), using the standard 70/15/15 train/val/test split. KhanQ (Gong et al., 2022) (1,034 paragraph–topic–question triples) serves as an English generalisation benchmark against the topic-QG literature. Chinese language left out due to limitations of the teacher model F.

Models. We evaluate the following model groups:

- **Teacher model:** Gemini 2.5 Flash (Comanici et al., 2025).
- **Student:** mT5-small (≈300M parameters).
- **Multilingual XQuAD baselines (zero-shot):** Aya Expanse 8B (Dang et al., 2024) and Qwen 2.5 7B (Qwen et al., 2025).
- **English KhanQ baselines (zero-shot):** Claude Sonnet 4.6, Claude Opus 4.5, Amazon Nova Pro/Lite/Micro, Llama 3.1 8B, Qwen 2.5 7B, DeepSeek 7B, Gemma2 9B, and Mistral 7B.
- **English reference model:** T5-small trained on SQuAD/MixSQuAD2X (Li et al., 2025).

Metrics. We report standard lexical metrics (F1, METEOR, ROUGE-L) as primary per-language overlap scores.

For **WikiSemRel** topic relevance, let E_q and E_r denote the Wikidata entity sets linked (via WAT) from q and r. We define:

$$\hat{q}_t = \begin{cases} 1, & \text{if } E_q \cap E_r \neq \varnothing, \\ \text{Jaccard}(E_q, E_r), & \text{otherwise.} \end{cases} \tag{3}$$

We also report the distractor-topic contrast (Li et al., 2025):

$$\Delta = \hat{q}_t - \hat{q}_{t'}, \tag{4}$$

where t' is a random in-paragraph distractor topic. Because WAT API entity linkers (Piccinno & Ferragina, 2014) have limited non-English coverage, WikiSemRel is measured on the English slice (XQuAD-en predictions for multilingual models, KhanQ for English-only models) as a semantic-fidelity proxy.

For answerability, we define **RQUGE-LLM** with the RQUGE architecture (Mohammadshahi et al., 2023):

$$\begin{aligned} a_c &= \text{QA}(q_c, D), \\ \kappa &= S(q_c, a_c, a_r, D), \end{aligned} \tag{5}$$

where the candidate question is generated as $q_c = \text{QG}(a_r, D)$ from context D and gold answer span a_r; $\text{QA}(\cdot)$ is instantiated as XLM-R to predict a_c from (q_c, D); and the scorer $S(\cdot)$ is instantiated as a Gemini 2.5 Flash judge to output acceptance score κ conditioned on (q_c, a_c, a_r, D). Methodology is described in Section E.3.

E. Extended Results and Analysis

E.1. Distillation vs. Bespoke Multilingual Training (RQ1–RQ3)

Table 3 presents results aggregated across XQuAD languages. We compare the three distilled students (MT5 GEMINI DISTIL 1X, 2X, and 3X), the same architecture trained on the bespoke augmented dataset (MT5 MIXXQUAD2X), and two open-weight zero-shot LLMs against the teacher.

Lexical overfitting in the bespoke baseline. The bespoke MIXXQUAD2X model achieves highest lexical overlap (F1 0.435) than either the teacher (F1 0.302) or the distilled student (F1 0.261). Taken alone, the result favours the bespoke dataset and augmentation approach of (Li et al., 2025). However, on WikiSemRel the same model drops to $\hat{q}_t = 0.269$ ($\Delta = 0.221$), below every LLM we tested and below the zero-shot open-weight 7B models on KhanQ (Table 4). The pattern is consistent with lexical overfitting: the model has learned the n-gram distribution of the small MixXQuAD2X training set

Table 3. QG performance on XQuAD test, macro-averaged over 11 languages. WSR columns are computed on the English slice: WSR $\hat{q}_t$ is the prescribed-topic score and WSR $\Delta = \hat{q}_t - \hat{q}_{t'}$ is the paper-faithful alt-topic-distractor difference (Li et al., 2025).

Model	Size	F1	METEOR	ROUGE-L	WSR $\hat{q}_t$	WSR Δ
Bespoke baseline						
mT5-small (MixXQuAD2X)	300M	**0.435**	**0.364**	**0.281**	0.269	0.221
Zero-shot LLMs						
Aya Expanse 8B	8B	0.303	0.251	0.180	–	–
Qwen 2.5 7B	7B	0.283	0.212	0.161	–	–
Distillation pipeline						
Gemini 2.5 Flash (teacher)	–	0.302	0.231	0.168	**0.675**	**0.605**
mT5-small (Gemini Distil 3x)	300M	0.261	0.186	0.143	0.584	0.518
mT5-small (Gemini Distil 2x)	300M	0.259	0.193	0.147	0.485	0.358
mT5-small (Gemini Distil 1x)	300M	0.266	0.192	0.145	0.334	0.239
Reference						
mT5-small (zero-shot, no fine-tune)	300M	0.024	0.013	0.022	–	–

rather than the underlying topic-to-question mapping. The model tracks the same lexical distribution as test set of XQuAD. We see similar lexical overfitting effect when optimising prompts for zero-shot LLMs Appendix K. The behaviour aligns with the finding of Nguyen et al. (Nguyen et al., 2024) that reference-based metrics on QG benchmarks reward surface similarity to the gold reference rather than question quality.

Teacher → student degradation (RQ2). Across the 11 evaluation languages, the distilled student retains 86% of the teacher's token F1 (0.261 vs. 0.302) and 85% of its ROUGE-L (0.143 vs. 0.168), despite having roughly 23× fewer active parameters and lower deployment cost. On WSR it retains $0.584/0.675 = 87\%$ of the teacher's $\hat{q}_t$ score (and $0.518/0.605 = 86\%$ of the WSR-diff), although the student is trained only on the teacher's outputs and not on its logits. These results address RQ1 and RQ2: distillation transfers multilingual QG capability into an SLM with limited degradation.

Lexical-vs-semantic trade-off under distillation (RQ3). Relative to the MIXXQUAD2X model, the distilled student trades lexical overlap for a 2.2× larger WSR ($\hat{q}_t = 0.584$ vs. 0.269; $\Delta = 0.518$ vs. 0.221). Because the distillation data is generated fresh per context at 3× scale, the student is exposed to a more diverse question distribution than the bespoke mixing approach can provide from XQuAD's 1,190 paragraphs per language. The distilled student's questions remain on-topic, whereas the bespoke model produces outputs that approximate n-gram averages of the training references.

E.2. Generalisation to English Educational Benchmarks

To check that multilingualism does not degrade English capability, we evaluated the same model family plus stronger LLMs on KhanQ (Table 4).

The bespoke English SOTA (Li et al., 2025) remains the top model on KhanQ, as expected: it is trained on SQuAD-derived augmentations that target this distribution. Two observations follow. First, the multilingual distilled mT5 (WSR $\hat{q}_t = 0.584$, $\Delta = 0.518$ on KhanQ; Table 3) tracks frontier English LLMs such as Claude Sonnet 4.6 ($\hat{q}_t = 0.585$, $\Delta = 0.502$). Second, the open-weight LLMs (Llama, Qwen, DeepSeek) and SQuAD-trained T5-small baselines fall in the 0.37–0.46 $\hat{q}_t$ range ($\Delta = 0.17$–0.34), below the bespoke SLM SOTA. The results suggest that WikiSemRel-measured topic control is acquired through training rather than emerging with scale.

E.3. Answerability vs. Topic Control (RQ4)

Lexical metrics are insufficient for multilingual evaluation: many valid semantic variations in Arabic, Russian, Thai, and Vietnamese yield near-zero token overlap despite being fluent, on-topic, and answerable from the source passage (see per-language breakdown in Appendix G). The original RQUGE (Mohammadshahi et al., 2023) is English-only because its backbone (a RoBERTa QA reader and a MOCHA span scorer) does not cover the other XQuAD languages. We therefore introduce **RQUGE-LLM**, a language-portable surrogate that preserves RQUGE's two-stage design:

Table 4. QG performance on the English KhanQ test set. WSR $\hat{q}_t$ is the prescribed-topic score; WSR $\Delta = \hat{q}_t - \hat{q}_{t'}$ is the alt-topic-distractor difference. The Li et al. row reports the published WSR-diff only.

Model	F1	METEOR	R-L	WSR $\hat{q}_t$	WSR Δ
Li et al. (Li et al., 2025) topicqg2x	0.321	0.220	0.216	0.735	**0.680**
Claude Sonnet 4.6 (opt)	0.272	0.205	0.172	**0.585**	**0.502**
Claude Opus 4.5 (opt)	0.281	0.216	0.180	0.555	0.458
Nova Lite	0.286	0.205	0.183	0.533	0.437
Llama 3.1 8B	0.251	0.197	0.161	0.422	0.272
Qwen 2.5 7B	0.251	0.195	0.163	0.405	0.243
DeepSeek 7B	0.238	0.196	0.156	0.404	0.172
T5-small (topic)	0.262	0.165	0.151	0.396	0.290
T5-small (baseline)	0.254	0.156	0.142	0.369	0.240
T5-base (zero-shot)	0.109	0.064	0.086	0.460	0.346
Gemini 2.5 Flash	0.275	0.210	0.178	**0.675**	**0.605**
mT5-small (3x)	0.262	0.172	0.156	0.584	0.518
mT5-small (2x)	0.273	0.189	0.189	0.485	0.358
mT5-small (1x)	0.258	0.165	0.153	0.334	0.239

- **Stage 1 (multilingual reader).** We replace the RoBERTa QA model with `deepset/xlm-roberta-large-squad2`, a multilingual XLM-R QA model that covers all 11 XQuAD languages. Given the generated question and the source passage, the reader extracts a predicted answer span.
- **Stage 2 (LLM judge).** We replace the MOCHA span scorer with Gemini 2.5 Flash acting as an LLM-as-judge with a MOCHA-style 1–5 rubric. The judge sees the passage, the generated question, the gold answer span, and the reader's extracted answer, and returns a single integer score in $[1, 5]$.

We score 179 (passage, question) pairs of XQuAD test set per model per language across the 11 languages. The RQUGE score requires reference answer span ground truth to measure the answerability.

Answerability ranks track the lexical pattern. Table 5 reports RQUGE-LLM macro-averaged across the 11 languages. The bespoke MIXXQUAD2X model ranks highest (2.496), followed by zero-shot Aya Expanse 8B (2.386) and the teacher LLM (2.111). The ranking appears to corroborate the lexical ranking from Table 3, but is consistent with the same overfitting interpretation. Because RQUGE-LLM's reader is a SQuAD-style extractive QA model, it tends to retrieve the gold span; MIXXQUAD2X is trained on the XQuAD reference questions whose answers are those gold spans, so its outputs are well-aligned with the reader's expected answers whereas LLM distilled questions target the general topic. The same model performs poorly on WSR (Table 3, $\hat{q}_t = 0.269$, $\Delta = 0.221$), which measures topic relevance rather than gold-span retrieval. The distilled student shows the opposite pattern: lower RQUGE-LLM (1.442) but a $2.2\times$ larger WSR ($\hat{q}_t = 0.584$, $\Delta = 0.518$), indicating questions that address the topic from formulations the reader retrieves less reliably.

The divergence is supported quantitatively by a cross-model correlation analysis on KhanQ (Appendix L): WSR and RQUGE-LLM are positively but only moderately correlated across 13 models (Pearson $r = 0.60$, $p = 0.029$; Spearman $\rho = 0.64$, $p = 0.019$), consistent with the two metrics capturing overlapping but distinct dimensions of question quality. The rank correlation explains only $\rho^2 \approx 41\%$ of model-rank variance, leaving roughly 59% unshared, enough to support the within-quality-band inversions Table 5 displays. WSR is far more tightly tied to the recall-weighted lexical family (F1 $\rho = 0.89$, ROUGE-L $\rho = 0.87$; both $p < 0.001$) than RQUGE-LLM is to any lexical metric.

A second, within-model line of evidence comes from prompt optimisation (Appendix K). Holding the LLM fixed and only changing the user prompt, the optimised variant raises Claude HAIKU's WSR by $+0.050$ (McNemar $p < 0.001$) while reducing its RQUGE-LLM by -0.081 (paired t, $p = 0.024$); on SONNET the same intervention moves WSR negligibly ($p = 0.73$) and reduces RQUGE marginally ($p = 0.07$). A within-model perturbation that improves topic grounding while degrading answerability is inconsistent with the two metrics measuring the same underlying signal. Combined with the cross-model correlation, the evidence addresses RQ4: WSR and RQUGE-LLM provide non-redundant signals about

Table 5. Reference-free metrics: RQUGE-LLM (answerability, 11-lang macro-avg) versus WSR (topic relevance, English KhanQ slice). WSR $\hat{q}_t$ is the prescribed-topic score and WSR Δ is the alt-topic-distractor difference. Higher is better for all three. Per-language RQUGE-LLM in Appendix H.

Model	RQUGE-LLM (1–5)	WSR $\hat{q}_t$	WSR Δ
mT5-small (zero-shot)	1.022	–	–
Bespoke baseline			
mT5-small (MixXQuAD2X)	**2.496**	0.269	0.221
Zero-shot LLMs			
Aya Expanse 8B	2.386	–	–
Qwen 2.5 7B	2.168	–	–
Distillation pipeline			
Gemini 2.5 Flash (teacher)	2.111	**0.675**	**0.605**
mT5-small (Gemini Distil 3x)	1.442	0.584	0.518
mT5-small (Gemini Distil 2x)	1.420	0.485	0.358
mT5-small (Gemini Distil 1x)	1.327	0.334	0.239

generated questions, and using either in isolation, or in place of the lexical metrics, would obscure the topic-control vs. answerability trade-off observed for the distilled student.

F. Additional Limitations

Scope of WikiSemRel. WAT (Piccinno & Ferragina, 2014) entity linkers we use have strong English coverage and weaker coverage in other XQuAD languages; we therefore report WSR ($\hat{q}_t$ and the WSR-diff Δ) as a semantic-fidelity signal on the English slice rather than per-language. A fully multilingual semantic-topic metric (e.g. multilingual Wikidata entity linking) is a natural extension.

Distillation technique. Our pipeline uses sequence-level data distillation: the teacher generates synthetic training pairs that the student learns from, without access to teacher logits or intermediate representations. More sophisticated schemes, such as logit-matching, intermediate-layer alignment (Jiao et al., 2020), or temporally adaptive interpolation (Shing et al., 2025), could reduce the teacher→student gap further. TAID (Shing et al., 2025) introduces a dynamic schedule that interpolates between the teacher and student output distributions during training, which may be especially beneficial for low-resource XQuAD languages where the synthetic data is sparser. We leave these extensions to future work.

Scope of RQUGE-LLM. Our multilingual RQUGE-LLM surrogate (Section E.3) reuses Gemini 2.5 Flash as the judge while the original RQUGE (Mohammadshahi et al., 2023) uses a smaller, frozen MOCHA (PLM) scorer. RQUGE-LLM scores are also biased toward gold-span retrieval, which favours bespoke models trained on XQuAD references; we therefore report it alongside, not instead of, WSR.

Chinese language exclusion. We exclude Chinese (zh) from the XQuAD experiments because the teacher model, Gemini 2.5 Flash, did not reliably produce usable distilled question generations for Chinese language in our pipeline within reasonable latency. One plausible explanation is language-specific distillation/safety constraints in the teacher model provider (Google) that are more restrictive for Chinese outputs. We therefore report results on the remaining 11 XQuAD languages.

G. Per-Language Performance

H. RQUGE-LLM Per-Language Scores

Table 7 reports RQUGE-LLM mean answerability scores (1–5) per language for all seven models, complementing the macro-averages in Table 5.

Table 6. Per-language F1 / ROUGE-L on XQuAD test ($n = 179$ per language). Avg row matches Table 3.

Language	mT5 (MixXQuAD2X) F1 / ROUGE-L	**mT5 (Distil 3x)** F1 / ROUGE-L	**mT5 (Distil 2x)** F1 / ROUGE-L	**mT5 (Distil 1x)** F1 / ROUGE-L	Gemini 2.5 Flash F1 / ROUGE-L
Arabic (ar)	0.356 / 0.045	0.148 / 0.018	0.158 / 0.023	0.163 / 0.008	0.204 / 0.020
German (de)	0.450 / 0.368	0.274 / 0.177	0.270 / 0.190	0.275 / 0.176	0.315 / 0.219
Greek (el)	0.421 / 0.127	0.262 / 0.052	0.254 / 0.055	0.274 / 0.061	0.319 / 0.093
English (en)	0.600 / 0.514	0.370 / 0.262	0.350 / 0.256	0.359 / 0.260	0.375 / 0.272
Spanish (es)	0.528 / 0.458	0.343 / 0.257	0.327 / 0.260	0.340 / 0.252	0.369 / 0.282
Hindi (hi)	0.451 / 0.080	0.305 / 0.033	0.313 / 0.030	0.303 / 0.028	0.359 / 0.035
Romanian (ro)	0.476 / 0.401	0.293 / 0.198	0.303 / 0.219	0.317 / 0.212	0.342 / 0.248
Russian (ru)	0.427 / 0.123	0.244 / 0.039	0.242 / 0.035	0.236 / 0.034	0.284 / 0.043
Thai (th)	0.170 / 0.124	0.074 / 0.017	0.072 / 0.019	0.081 / 0.030	0.079 / 0.024
Turkish (tr)	0.449 / 0.398	0.245 / 0.186	0.241 / 0.191	0.256 / 0.195	0.278 / 0.228
Vietnamese (vi)	0.459 / 0.459	0.309 / 0.330	0.317 / 0.341	0.320 / 0.337	0.393 / 0.380
Avg (11 langs)	**0.435 / 0.281**	**0.261 / 0.143**	**0.259 / 0.147**	**0.266 / 0.145**	**0.302 / 0.168**

The gap between high-resource Indo-European languages (en, es, de) and morphologically rich or non-space-tokenised ones (th, ar, hi) highlights the tokenisation and morphological challenges inherent to multilingual QG.

Table 7. Per-language RQUGE-LLM mean score (1–5). Avg row matches Table 5.

Language	mT5 (zero)	mT5 (Mix2X)	**mT5 (Distil 3x)**	**mT5 (Distil 2x)**	**mT5 (Distil 1x)**	Gemini (T)	Aya 8B	Qwen 7B
Arabic (ar)	1.000	2.469	1.525	1.385	1.374	2.408	**2.816**	2.240
German (de)	1.000	2.559	1.425	1.358	1.313	2.061	**2.587**	2.307
Greek (el)	1.045	**2.514**	1.335	1.385	1.162	2.045	2.391	1.922
English (en)	1.056	**2.726**	1.559	1.531	1.391	1.927	2.503	2.715
Spanish (es)	1.050	**2.894**	1.615	1.525	1.475	2.279	2.665	2.408
Hindi (hi)	1.034	**2.235**	1.385	1.397	1.251	2.112	2.190	1.760
Romanian (ro)	1.000	2.441	1.447	1.469	1.296	2.227	**2.570**	2.173
Russian (ru)	1.028	2.413	1.441	1.475	1.235	2.128	**2.760**	2.402
Thai (th)	1.000	**2.765**	1.520	1.503	1.542	2.056	1.436	2.045
Turkish (tr)	1.034	**2.184**	1.380	1.313	1.285	1.810	2.073	1.749
Vietnamese (vi)	1.000	**2.257**	1.235	1.274	1.274	2.162	**2.257**	2.128
Avg (11 langs)	**1.022**	**2.496**	**1.442**	**1.420**	**1.327**	**2.111**	**2.386**	**2.168**

Bold marks the per-language top score. MIXXQUAD2X ranks first in 7/11 languages and Aya Expanse 8B in 5/11 (with one tie); The distilled GEMINI DISTIL students (1x, 2x, and 3x) trail the bespoke and zero-shot LLM ranks consistently. The answerability signal complements the lexical overfitting in Table 6 and the WSR degradation for MIXXQUAD2X in Table 3.

I. Implementation Details and Hyperparameters

All hyperparameters below are the ones actually used, read verbatim from `config/pipeline.yaml` in the released code.

Student input/target format. All students (GEMINI DISTIL 1x, 2x, and 3x, and MIXXQUAD2X) share the same token format:

`<topic> {T} <context> {C}` → `{Q}`

where `<topic>` and `<context>` are special tokens added to the mT5 vocabulary, T is the target topic, C the source paragraph, and Q the gold question.

Student training (mT5). The student is initialised from the pre-trained `google/mt5-small` checkpoint (≈300M parameters). Optimiser: AdamW with a linear warmup of 100 steps. Learning rate: $3e{-4}$. Per-device batch size: 32, gradient accumulation steps: 2 (effective batch size 64). Weight decay: 0.01. Maximum input length: 200 tokens; maximum target length: 45 tokens. Training runs for up to 20 epochs with early stopping on validation loss (patience 3) to avoid overfitting to the synthetic generations. No mixed precision is used by default (we observed training instability with `fp16` on mT5-small).

Bespoke baseline training. MIXXQUAD2X shares the above configuration but is trained on the Li et al. (Li et al., 2025) augmented mixing dataset (MIXXQUAD2X) derived from XQuAD's train split.

Data splits. For each XQuAD language we use the official 70/15/15 train/val/test split (≈1190 paragraphs, stratified by topic). The distillation pipeline operates only on train+val; XQuAD-test is untouched and constitutes our held-out evaluation set.

Generation and evaluation. At inference we use beam search with 10 beams and length penalty 1.0, keeping the top-1 sequence. Lexical metrics are F1 (SQuAD-style), NLTK METEOR, and ROUGE-L via `rouge-score`. WikiSemRel (Li et al., 2025) uses WAT API (Piccinno & Ferragina, 2014) for entity linking and a random in-paragraph topic as the noisy context; we report the mean $t - t'$ difference where t is relatedness to the target topic and t' to the noisy context.

Hardware. All student training and evaluation was performed on a single NVIDIA Laptop RTX 4060 (8 GB) GPU.

J. Knowledge Distillation Prompts

The $3\times$ synthetic distillation dataset is produced by Gemini 2.5 Flash in two stages: topic extraction, then per-topic question generation. Both prompts enforce the paragraph's source language so that the student learns to generate in the target language rather than in translation. Decoding parameters are `temperature = 0.4` and `max_output_tokens = 1024` for both stages. Contexts are truncated to 2,000 characters before prompting. The prompts below are reproduced verbatim from `scripts/generate_xquad_distillation.py`.

Stage 1: Topic extraction prompt.

> *Read the following paragraph and identify the {k} most important topics or key concepts that could be asked about as questions.*
>
> *Paragraph: {context}*
>
> *Return ONLY a JSON array of {k} short topic phrases (not full sentences), in the same language as the paragraph. Example format:* `["topic one", "topic two", "topic three"]`
> *Return nothing else.*

where {k} is fixed to 3 in all our experiments and {context} is the XQuAD paragraph. The response is parsed as JSON with a permissive fallback to newline-separated / numbered-list formats, and topics are filtered to reject JSON artefacts and out-of-range lengths (2–80 characters).

Stage 2: Per-topic question prompt.

> *Generate a question about the given topic based on the paragraph.*
> *Respond in {lang_name}, the same language as the paragraph.*
>
> *Paragraph: {context}*
>
> *Topic: {topic}*
>
> *Requirements:*
>
> - *Generate ONE clear question about the topic*
> - *Write the question in {lang_name}*
> - *The question must be answerable from the paragraph*
> - *Length: 8–15 words*
> - *Output ONLY the question text*

where {lang_name} is the full English name of the target language (Arabic, German, Greek, . . . , Chinese) rather than its ISO code, which we found improves non-English fluency. Generated questions beginning with the literal prefix "`Question:`" have it stripped. All model outputs are cached per (context, topic) pair so that the pipeline can be resumed incrementally if upstream rate limits or errors interrupt generation.

Table 8. Prompt optimisation on KhanQ. **Top:** lexical metrics. **Bottom:** WikiSemRel (English, WAT entity linking).

Stage 1: lexical (`optimised` − default)					
Model	n_{opt}	ΔB1	ΔB4	ΔF1	ΔRL
Haiku	653	**+.006**	−.003	**+.005**	**+.008**
Sonnet	653	**+.003**	−.005	.000	**+.002**
Opus	588	**+.009**	−.004	**+.005**	**+.007**
Stage 2: topic grounding (KhanQ-en WSR)					
Model	n	WSR base	WSR `optimised`	ΔWSR	Δexact
Haiku	640	0.520	0.570	**+0.050**	**+35**
Sonnet	640	0.592	0.595	**+0.003**	**+4**

K. Effect of Prompt Optimisation: Lexical, then Topic Grounding

We document the effect of a single prompt-engineering pass on zero-shot baselines as a two-stage progression: prompt tuning produces modest *lexical* gains (BLEU-1, F1, ROUGE-L) consistently, and reveals a *topic-grounding* effect (WSR) on KhanQ that the lexical metrics underestimate, most visibly on the smaller HAIKU model, while the stronger SONNET is already near the WSR ceiling for this prompt family.

Prompt variants. The default prompt asks for a single concise question with a length and starter-word constraint. The optimised prompt adds (i) an explicit requirement to mention the topic by name (or a close synonym), (ii) a tighter 10–15-word length, (iii) four in-context exemplars, and (iv) explicit guidance to ask about mechanisms, relationships, comparisons, or definitions.

Stage 1: lexical. The optimised prompt produces small but consistent gains in recall-weighted lexical metrics across all three Claude tiers (Table 8, top half): BLEU-1 by 0.003–0.009 absolute (2–6 % relative), F1 by 0.000–0.005, ROUGE-L by 0.002–0.008. BLEU-4 moves slightly the other way (-0.003 to -0.005), consistent with optimised prompt licensing more diverse 4-gram phrasing rather than template-style copying.

Stage 2: topic grounding. The same prompt change moves WikiSemRel (Table 8, bottom half), measured on KhanQ-en using WAT entity linking(Piccinno & Ferragina, 2014). The effect is, however, strongly model-dependent. HAIKU's WSR rises from 0.520 to 0.570 ($+0.050$ absolute, $+9.6\,\%$ relative). SONNET only moves from 0.592 to 0.595 ($+0.5\,\%$ relative): its baseline WSR is already close to the optimised HAIKU number.

These results are consistent with surface n-gram overlap and entity-level topic grounding being separable dimensions of question quality (Section E.3). Adding "mention the topic by name" to the prompt steers the generator toward entities the WSR scorer recognises. The smaller HAIKU shows roughly a $2\times$ amplification of the lexical signal in WSR space, whereas SONNET, whose baseline is already near the WSR ceiling for this prompt family, captures most of the available improvement without prompt engineering.

L. Relationship Between WSR and RQUGE-LLM

We test how WSR and RQUGE-LLM relate to one another and to the lexical metrics. The analysis is run across the KhanQ evaluation suite of 13 models (zero-shot LLMs Claude HAIKU/SONNET/OPUS with default and `:opt` prompts where available, Amazon Nova PRO/LITE/MICRO, English-T5 SQuAD-baseline and SQuAD-topic, and three mT5 students: MIXXQUAD2X, GEMINI DISTIL 1X, 3X). We exclude both zero-shot reference checkpoints (English T5-base and mT5-small): their outputs score at the floor of every metric simultaneously and would dominate the Pearson estimates without informing the underlying relationship. The WSR field is the $\hat{q}_t$ WAT-jaccard score and RQUGE-LLM is the per-model macro-average from Section E.3.

WSR and RQUGE-LLM are positively but moderately correlated. The two metrics correlate at Pearson $r = 0.60$ ($p = 0.029$) and Spearman $\rho = 0.64$ ($p = 0.019$): both significant, but the rank correlation accounts for only about 41% of model-rank variance (ρ^2). The remaining variance is consistent with the within-quality-band rank inversions noted in the main paper, e.g. MIXXQUAD2X's RQUGE-LLM lead and lower WSR (Table 5), or the distilled student's RQUGE deficit

Table 9. Correlation between metrics on KhanQ ($n = 13$ models, both zero-shots dropped). Two-tailed p-values for H_0: $\rho = 0$. **Bold** marks $p < 0.05$.

Metric A	Metric B	Pearson r	p_r	Spearman ρ	p_ρ
WSR	RQUGE-LLM	**0.60**	**0.029**	**0.64**	**0.019**
WSR	BLEU-1	**0.64**	**0.018**	**0.72**	**0.006**
WSR	BLEU-4	−0.00	0.999	−0.30	0.325
WSR	F1	**0.86**	**0.000**	**0.89**	**0.000**
WSR	ROUGE-L	**0.90**	**0.000**	**0.87**	**0.000**
RQUGE-LLM	BLEU-1	0.21	0.486	0.26	0.394
RQUGE-LLM	BLEU-4	−0.16	0.602	−0.34	0.255
RQUGE-LLM	F1	0.47	0.103	**0.62**	**0.025**
RQUGE-LLM	ROUGE-L	0.44	0.130	0.45	0.122

together with a $2.2\times$ larger WSR. The two metrics therefore overlap but capture distinct signals: a strong model tends to score well on both, but a high score on one does not imply a high score on the other.

WSR correlates strongly with recall-weighted lexical metrics. WSR ranks models almost identically to ROUGE-L ($r = 0.90$, $\rho = 0.87$; both $p < 0.001$) and to F1 ($r = 0.86$, $\rho = 0.89$; both $p < 0.001$). Both lexical metrics are recall- and unigram-weighted, and a question that hits the same Wikipedia entity as the reference will typically share content words with it. WSR is at the same time *uncorrelated* with BLEU-4 ($r \approx 0.0$, $p = 0.999$; $\rho = -0.30$, $p = 0.33$): producing the right topic entity is essentially orthogonal to producing matching 4-gram phrases.

RQUGE-LLM tracks lexical metrics far more weakly. RQUGE-LLM has only one significant correlation in this set, a moderate Spearman with F1 ($\rho = 0.62$, $p = 0.025$). All Pearson correlations with lexical metrics are non-significant ($|r| \leq 0.47$, $p \geq 0.10$). The answerability score is closer to a distinct dimension than to a re-encoding of n-gram overlap, and most of the WSR $\leftrightarrow$ RQUGE correlation reported above is mediated by the component of WSR that aligns with recall-weighted lexical features rather than by an independent RQUGE–lexical relationship.