FINEWEB2-HQ-QA: NATIVE DOCUMENT-GROUNDED QA FOR MULTILINGUAL PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in data curation have substantially improved the quality of multilingual web corpora, yet large-scale synthetic resources for non-English pre-training remain scarce and typically rely on machine translation from English-source data. We introduce FINEWEB2-HQ-QA, a multilingual dataset containing 193B tokens of source documents and generated QA across 20 languages. We generate QA pairs directly in each target language from native FineWeb2-HQ web documents rather than from English-source text (*native QA*). This approach is designed to reduce dependence on English-source topic distributions by grounding generation in target-language web content. At 3B and 8B scales, we compare our native document-grounded QA against non-QA baselines and Nemotron's English and machine-translated QA. Relative to matched non-QA baselines, our dataset improves Global and Local Knowledge across all four training settings while largely preserving or improving aggregate general language understanding. Among individual QA sources, translated QA yields the largest observed Global Knowledge gains across all four settings, whereas native QA consistently shows higher observed scores than translated QA on general language understanding and achieves strong Local Knowledge gains at 8B. In natural-mixture settings, combining sources yields a more balanced profile across categories without consistently exceeding the best individual source per category. We release our dataset alongside our generation and filtering pipelines to support future work on multilingual synthetic pre-training data.

1 INTRODUCTION

Model-based quality filtering of raw web text has become an established standard for curating large-scale pre-training corpora (Penedo et al., 2024; Li et al., 2024), a foundation recently extended to multilingual settings with the release of FineWeb2-HQ (Messmer et al., 2025). In parallel, the field is increasingly shifting toward synthetic data to further augment pre-training corpora (Eldan & Li, 2023; Gunasekar et al., 2023; PleIAs, 2025; Niklaus et al., 2026; Su et al., 2025; Liu et al., 2025). Within this space, we observe a growing trend toward generating question-and-answer (QA) augmentations directly from high-quality web text (Su et al., 2025; Liu et al., 2025) or incorporating existing QA datasets (Olmo et al., 2025).

Despite this shift toward synthetic QA augmentation, most efforts remain heavily English-centric (Eldan & Li, 2023; Gunasekar et al., 2023; PleIAs, 2025; Niklaus et al., 2026; Su et al., 2025; Liu et al., 2025). For non-English languages, synthetic datasets are often translated from English sources, such as the Translated Diverse Question Answer (TDQA) subset in Nemotron-CC-v2 (Basant et al., 2025). Beyond potential translationese artifacts (Vanmassenhove et al., 2021), translated synthetic data inherits the topical distribution and knowledge focus of their English sources, potentially underrepresenting knowledge and domains specific to target-language communities. This motivates alternatives to English-centric sourcing, alongside broader concerns about cultural homogenization and the loss of linguistic diversity (Nguyen et al., 2024). Generating from target-language documents aims to provide an alternative source distribution, although language-based sourcing does not guarantee geographic, cultural, or dialectal representativeness or that every source document was originally authored in that language.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Motivated by recent work on large-scale multilingual curation (Carranza et al., 2026), we introduce FINEWEB2-HQ-QA, a multilingual synthetic complement to FineWeb2-HQ for all 20 available languages. We prompt a large language model (LLM) to generate QA pairs directly from FineWeb2-HQ documents in each target language, rather than translating English-source QA. By grounding generation in target-language web documents, we base the QA content on the topics covered by those documents. We use *native QA* to denote this target-language sourcing and generation procedure.

We support the choice of QA augmentation through cooldown ablations over publicly released Nemotron-CC-v2 subsets (Basant et al., 2025), following established methodology for comparing pre-training data mixtures (Grattafiori et al., 2024; Apertus et al., 2025). In our 3B setup, English and translated QA yield stronger aggregate results than the high-quality web and synthetic rephrasing subsets. Benefits from QA augmentation also persist when cooling down an Apertus 8B model from a checkpoint trained on 9.5T tokens (Apertus et al., 2025). These supporting experiments are reported in Appendix C.

We therefore ask: *How does QA generated from target-language FineWeb2-HQ documents compare with publicly released English-source QA and its translations for multilingual pre-training?* Because these resources differ in source documents, document selection, generation pipelines, and filtering, our experiments compare complete resources rather than isolate the causal effect of native generation versus translation. Across the evaluated settings, translated QA shows the highest observed global-knowledge scores among single sources, while native QA shows higher observed GLU Avg than translated QA in all four settings and strong observed Local gain at 8B. Combining these resources changes the balance across evaluation categories, but does not uniformly outperform the strongest individual source.

Our contributions are as follows:

- We introduce FINEWEB2-HQ-QA, a 193B-token[1] native document-grounded multilingual QA augmented dataset covering 20 languages, together with the generation and filtering pipelines.[2]
- We compare individual and combined native, English, and translated QA sources under matched token budgets, including a shared-language configuration that equalizes token allocation across the 13 languages shared by the two multilingual QA resources.
- We observe distinct performance profiles and mixing trade-offs under fixed budgets to guide practical mixtures. Among individual QA sources, translated QA shows the largest observed Global Knowledge gains, while native QA shows competitive observed GLU and strong observed Local gains at 8B. We further characterize how combining these sources changes the balance across evaluation categories under a fixed token budget.

2 RELATED WORK

Synthetic Data in LLM pre-training. While the performance of LLMs was initially improved by scaling the amount of training data (Brown et al., 2020), increasing attention has been paid towards better-curated datasets (Dodge et al., 2021; Penedo et al., 2023; 2024; Li et al., 2024; Raffel et al., 2020). More recently, the focus has moved towards synthetic data generation. TinyStories (Eldan & Li, 2023) was a pioneering work in this area showing that models with fewer than 10 million parameters could learn coherent English solely from synthetic text, while Gunasekar et al. (2023) demonstrated that using synthetic "textbook-quality" data can achieve competitive performance. Building on these foundations, Cosmopedia (Allal et al., 2024) was released as a large-scale, open-source synthetic dataset. The release of the open-source pleias/SYNTH dataset (PleIAs, 2025) has demonstrated how high-quality synthetic derivatives can enable small-scale reasoning models to achieve strong performance for their parameter count. The reliance on synthetic data in recent open-source open-data LLMs (Basant et al., 2025; Blakeman et al., 2025; Liu et al., 2025; Olmo et al., 2025) confirms this trend, alongside Kang et al. (2025) showing that mixing around 30% of

[1] Measured using the Apertus-70B-Instruct-2509 tokenizer.

[2] An anonymized preview (1k samples) with generation and filtering code is provided as supplementary material for review. Upon acceptance, we will release the unfiltered generated dataset and the generation and filtering pipelines. The reported 193B-token size refers to the retained source-document–QA samples after filtering.

synthetic data with natural web data accelerates pre-training. The work by Niklaus et al. (2026), exploring English-based synthetic generation, identifies QA pairs derived from web text among the most effective augmentation formats. This aligns with Allen-Zhu & Li (2024), who highlight the importance of knowledge-augmentation in the pre-training stage.

Multilingual Web Data Curation. With most open-source open-data LLMs focusing primarily on English, multilingual corpora curation has historically lagged behind. Several efforts have attempted to close this gap (Penedo et al., 2025; De Gibert et al., 2024). The emergence of massively multilingual models such as Apertus (Apertus et al., 2025) makes curating multilingual text resources increasingly important. FineWeb2-HQ (Messmer et al., 2025) addressed this gap by applying model-based filtering to multilingual web text. More recently, Carranza et al. (2026) emphasized the importance of natively multilingual curation at the 20-trillion-token scale.

Multilingual Synthetic Augmentation. Several large-scale multilingual synthetic resources rely on machine translation. Datasets such as the *Translated Diverse Question Answer (TDQA)* subset in Nemotron-CC-v2 (Basant et al., 2025), covering 15 languages, and MT-Nemotron (MultiSynt, 2025), which translates a subset of the English Nemotron-CC dataset into multiple languages, provide broad multilingual coverage but risk introducing translationese artifacts (Vanmassenhove et al., 2021) and inherit the topical distribution and knowledge focus of their English sources. Some efforts incorporate language-specific data more directly: MURI-IT (Köksal et al., 2025) generates instruction-output pairs from existing human-written texts across 200 languages via reverse instructions, using existing target-language texts rather than relying solely on translation from English for post-training. SeaLLMs (Nguyen et al., 2024; Ng et al., 2025) curate region-specific corpora for Southeast Asian languages for continual pre-training and post-training. More directly, Nemotron Nano 2 (Basant et al., 2025) also explores multilingual QA generated from target-language Wikipedia articles alongside translated English QA. Our work extends this direction with a 193B-token, 20-language dataset of source documents and QA pairs grounded in FineWeb2-HQ web documents, along with resource-level profiling against publicly released English and translated QA under two training regimes and two language-mixture configurations. We also make our generation and filtering pipelines public for the community to build upon.

3 The FineWeb2-HQ-QA Dataset

This section describes the construction of FineWeb2-HQ-QA, including native-language QA generation, quality filtering, and the resulting dataset and release.

3.1 Dataset Generation

Motivated by the growing trend of generating question-and-answer (QA) augmentations directly from high-quality web text during pre-training (Su et al., 2025; Liu et al., 2025; Olmo et al., 2025) and our supporting Nemotron subset ablations (Appendix C), we select QA pair augmentation as our synthetic data strategy. Below, we describe the generation setup and the four-stage filtering pipeline we apply to improve dataset quality.

QA Augmentation. We adopt the QA generation technique from Nemotron-CC (Su et al., 2025) (the base for Nemotron-CC-v2 (Basant et al., 2025)). We use the same prompt template designed for Nemotron's English QA generation but adjust it to prompt the LLM to directly generate the QA pairs in the respective target language. The prompt is shown in Appendix D.1. As source text, we sample documents from FineWeb2-HQ (Messmer et al., 2025). We generate the QA pairs using Qwen3-30B-A3B (Yang et al., 2025), the same model used to translate the Nemotron-CC-v2 DQA samples. This matches the model family used in the TDQA translation stage, but does not equalize the complete generation pipelines. We leave reasoning enabled, restrict output length to 8192 tokens, and remove the reasoning trace for the final dataset. Each training sample concatenates the source document with its generated QA pairs.

3.2 FILTERING.

To improve quality, we apply four sequential filters to the generated QA pairs. First, we remove aborted reasoning traces by requiring a completed `</think>` tag. Second, we filter language mismatches of the QAs using GlotLID (Kargaran et al., 2023), following FineWeb2 (Penedo et al., 2025), and retain samples with a language-identification score at least 0.9. Third, we remove degenerate generations with the Gopher repetition filter (Rae et al., 2021), targeting repeated QAs and phrases. Finally, we apply 13-gram benchmark decontamination using `datatrove` pipeline adapted from `lighteval`. Filter definitions, threshold selection, per-language removal rates, and removed examples are provided in Appendix D.

3.3 DATASET OVERVIEW AND RELEASE

The final dataset spans 20 languages, with a total of 188.8 million (M) raw web documents and 174M retained document-grounded QA samples. The retained dataset contains 193B tokens (document + QA), measured with the Apertus-70B-Instruct-2509 tokenizer. Per-language statistics are provided in Table 1.

Automated quality audit. We use Gemini 3.6 Flash to audit 1,500 randomly sampled documents across six languages spanning diverse scripts and language families[3] (250 documents per language). The judge assesses each QA pair against its source passage using a fixed rubric covering seven issue types: duplicate, malformed, false premise, trivial, unanswerable, unfaithful, and wrong language. It also flags source spam, suspected source factual errors, toxicity in the source and generated QA, and language mismatches. Unfaithful answers and unanswerable questions are the most frequently flagged QA issues, affecting 5.3–7.0% and 2.1–3.5% of pairs per language, respectively (categories may overlap). Spam is the most frequent document-level flag, reaching 6.8% in Chinese and 5.2% in Indonesian, while suspected factual errors are flagged in 1.6–3.2% of source documents per language. These are automated assessments rather than human-validated error rates: QA faithfulness is judged against the source, while suspected factual errors rely on the judge's own knowledge. Repeated runs show variability, particularly for rare or subjective categories, so these rates should be interpreted as approximate. The full rubric, per-language results, and repeatability and error-injection tests are reported in Appendix F.

Release. We release the raw version of the dataset together with our filtering code, enabling the community to refine the filtering approach, tailor it to specific languages and scripts, or filter individual QA pairs.

4 EXPERIMENTAL SETUP

4.1 MODELS AND TRAINING DESIGN

To evaluate the expected impact of our data subsets on pre-training data mixtures, we perform cooldown ablations following established methodology (Grattafiori et al., 2024; Apertus et al., 2025). A cooldown ablation continues training from a fixed checkpoint while annealing the learning rate, allowing alternative data mixtures to be compared under a fixed additional training budget. We construct a 100B-token training pool consisting of 70B tokens from a base mixture (detailed in Appendix B; Table 2) and 30B tokens from the target data subset. This 30% synthetic ratio is consistent with the findings of (Kang et al., 2025). From this pool, we uniformly sample tokens for training, consuming approximately 50B tokens for our 3B-parameter runs and 63B tokens for our 8B-parameter runs. In both cases, we anneal the learning rate using a `1-sqrt` decay schedule (Hägele et al., 2024), reducing it either to 0 or to 10% of the initial learning rate. We apply this methodology across both model scales:

- **3B Base Model:** We pre-train a 3B parameter model on 200B tokens using a base mixture of 85% FineWeb and 15% top-33% FineWeb2-HQ from Apertus Phase 1 (Apertus et al., 2025). This is intentionally a weak checkpoint. By starting with a lower-quality model, we amplify the signal and

[3] Arabic, Chinese, German, French, Indonesian, and Japanese.

Language	Input Docs	ART (%)	LID (%)	DS (%)	BC (%)	TD (%)	FD	TT (B)
arb_Arab	3,353,745	2.14	16.96	0.26	0.01	19.38	2,703,861	2.4
ces_Latn	2,452,481	3.80	0.04	0.10	0.00	3.94	2,355,940	2.9
cmn_Hani	26,076,021	0.60	5.13	10.76	0.40	16.89	21,671,485	33.9
dan_Latn	3,021,637	2.67	0.32	0.16	0.00	3.15	2,926,495	3.2
deu_Latn	20,759,205	1.59	0.75	0.04	0.03	2.41	20,259,211	20.5
ell_Grek	1,619,179	7.64	0.04	1.89	0.01	9.59	1,463,951	1.6
fas_Arab	1,984,552	1.92	0.31	0.74	0.03	3.00	1,925,037	1.9
fra_Latn	15,973,025	1.71	0.66	0.13	0.05	2.55	15,565,043	15.1
hun_Latn	2,166,540	3.63	0.03	0.19	0.00	3.84	2,083,291	2.5
ind_Latn	3,378,722	1.80	0.37	0.17	0.06	2.40	3,297,569	3.8
ita_Latn	10,195,425	1.29	0.52	0.16	0.02	2.00	9,991,872	9.3
jpn_Jpan	15,043,155	0.59	0.02	37.68	0.11	38.40	9,266,235	11.1
nld_Latn	6,680,086	1.34	0.65	0.11	0.03	2.13	6,537,938	6.1
pol_Latn	6,640,453	1.34	0.01	0.14	0.03	1.52	6,539,393	6.9
por_Latn	7,848,376	1.00	0.09	0.42	0.10	1.60	7,722,466	7.1
rus_Cyrl	34,829,012	4.21	0.06	0.12	0.01	4.40	33,297,435	38.4
spa_Latn	17,346,026	0.93	0.05	0.27	0.04	1.29	17,122,966	16.3
swe_Latn	3,306,324	1.82	0.01	0.13	0.00	1.96	3,241,390	3.4
tur_Latn	2,796,347	1.06	0.03	0.51	0.05	1.66	2,749,929	2.8
vie_Latn	3,355,737	0.90	0.01	0.28	0.12	1.31	3,311,940	4.0
Overall	**188,826,048**	**1.87**	**1.24**	**4.64**	**0.09**	**7.83**	**174,033,447**	**193.2**

Table 1: Web document retention and removal rates by language. Drop percentages are relative to initial input. ART, LID, DS, and BC denote *Aborted Reasoning Traces*, *Language Mismatch*, *Degenerate Samples*, and *Benchmark Contamination* filters, respectively. TD, FD, and TT denote total drop, final document count, and total token count, respectively.

make it easier to observe the gains from the high-quality data during cooldown. This explores the effect of using our dataset in a highly parameter and compute constrained setup.

- **8B Base Model:** To verify that our findings hold at scale, we use an intermediate Apertus 8B checkpoint already trained on 9.5T tokens as the starting point of our learning rate cooldown. This explores the effect of introducing our dataset late in training, where the added QA component accounts for less than 1% of total tokens seen.

The full training setup, shared across all experiments, is detailed in Appendix B.

4.2 Evaluation.

We conduct all evaluations using LM-Harness (Gao et al., 2024) across a diverse set of tasks. To establish English performance baselines, we include MMLU (Hendrycks et al., 2021b;a), OpenBookQA (Mihaylov et al., 2018), SQuAD v2 (Rajpurkar et al., 2018; 2016), Lambada OpenAI (Radford et al., 2019; Paperno et al., 2016), ARC Easy and Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), WinoGrande (Keisuke et al., 2019), PIQA (Bisk et al., 2020), and CommonsenseQA (Talmor et al., 2019). Reported standard errors reflect evaluation-set sampling, assuming independence across benchmarks and clustering translations of the same item across languages within translated benchmarks. Since each configuration was trained once, they do not capture run-to-run training variance.

Following Apertus (Apertus et al., 2025), we group multilingual benchmarks into three skill categories: *Global Knowledge*, *Local Knowledge*, and *General Language Understanding (GLU)*, respectively evaluated using GlobalMMLU (Singh et al., 2025); INCLUDE (Romanou et al., 2024); and XNLI (Conneau et al., 2018), PAWS-X (Yang et al., 2019), XCOPA (Ponti et al., 2020), XWinograd (Muennighoff et al., 2022; Tikhonov & Ryabinin, 2021), m_arc and m_hellaswag (Dac Lai et al., 2023). For each benchmark, we use the default LM-Harness metric and macro-average over its covered languages. Our main category scores then average these benchmark-level scores. Per-language aggregates and their aggregation procedure are provided in Appendix E.2. Note that m_arc and m_hellaswag are evaluated exclusively for the 8B model, as the 3B model exhibits near-random performance across all languages on these tasks. While we also evaluated CulturalBench (Chiu et al.,

2024), we found macro-averaged results highly volatile due to small per-country sample sizes (as few as 7 questions), and therefore report these results in Appendix E.3 alongside micro-averaged scores.

5 Results and Discussion

We evaluate whether our native document-grounded QA provides useful pre-training signal beyond existing publicly released English and translated QA datasets. Figure 1 summarizes gains over matched non-QA cooldown baselines across three category-level metrics: *Global Knowledge*, *Local Knowledge*, and average *General Language Understanding* (GLU).

Specifically, we ask whether (Q1) QA augmentation of FineWeb2-HQ (FineWeb2-HQ-QA) outperforms the baseline; (Q2) what performance profiles native, translated (TDQA) and English (DQA) QA resources show under matched budgets; and (Q3) how these datasets perform individually and in combination under a realistic language distribution.

We address these questions using two distinct data mixture configurations to account for variations in language coverage, detailed below. The ablations follow the general experimental design established in Section 4.

5.1 Data Mixture Configurations.

Because FineWeb2-HQ-QA and Translated Diverse QA (TDQA) vary in their language distribution, we evaluate our native QA augmentation using the two setups below:

Balanced Mixture (BM): This setup equalizes the target-resource allocation across the 13 languages shared between FineWeb2-HQ-QA and TDQA[4]. The 30B-token target component is divided equally across these languages, while the 70B-token base mixture is unchanged across conditions. When multiple datasets are combined, the budget is first split equally among them (e.g., 15B tokens each when two datasets are used), and each resulting share is then distributed evenly across the languages. In the baseline, the 30B-token target component is FineWeb2-HQ-QA documents with QA pairs removed, subsampled equally across the same 13 languages. This comparison evaluates replacing part of a fixed document-only token budget with document-grounded QA, using the same underlying source-document collection. We report multilingual results only for these 13 languages, intersected with each benchmark's coverage.

Natural Mixture (NM): This configuration reflects a more realistic, unconstrained training environment. When multiple datasets are combined, the target-resource budget is first split equally among them (e.g., divided by 2 or by 3), and each share is then distributed across languages according to the dataset's natural proportions. Notably, it expands the scope to the full set of 20 (or 15) available languages in FineWeb2-HQ-QA (TDQA) and evaluates models on the complete benchmarks across all offered languages. This allows us to observe the aggregate benefits closer to the expected effect in practical applications. For the NM baseline, the 30B-token target component consists of FineWeb2-HQ-QA documents with the QA pairs removed, while the 70B-token base mixture remains unchanged.

5.2 Performance impact of Native QA Augmentation (Q1)

We first ask whether FineWeb2-HQ-QA improves over the non-QA cooldown baseline. Across both training regimes and mixture configurations, native QA scores above the baseline in both knowledge categories. At 3B, FineWeb2-HQ-QA improves Global Knowledge by 4.9 points in BM and 5.6 points in NM, while improving Local Knowledge by 2.4 and 1.5 points respectively. At 8B, the gains remain positive despite starting from a much stronger checkpoint already trained on 9.5T tokens: native QA improves Global Knowledge by 2.7 points in BM and 2.6 points in NM, and improves Local Knowledge by 3.6 and 3.4 points respectively. These results suggest that document-grounded QA is an effective format for making factual knowledge accessible to the model, consistent with

[4]The 13 languages are: Arabic, Chinese, Danish, German, French, Italian, Japanese, Dutch, Polish, Portuguese, Russian, Spanish, and Swedish.

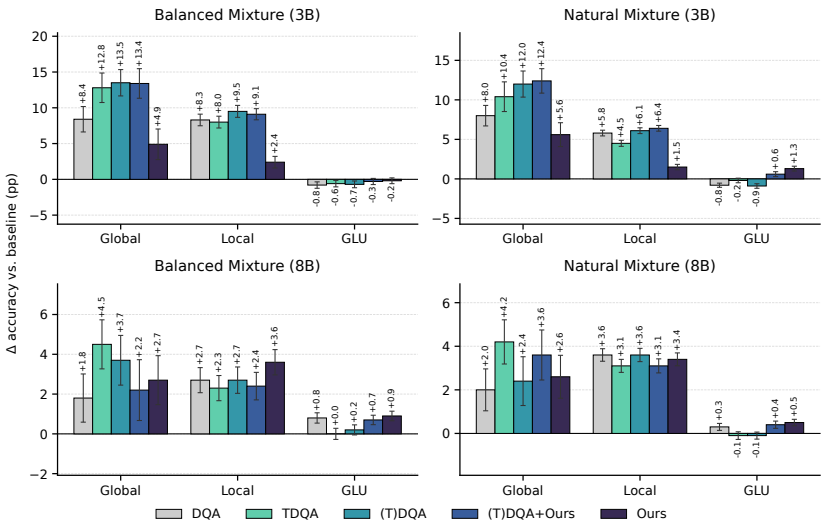


Figure 1: Category-level gains over matched non-QA cooldown baselines for Global Knowledge, Local Knowledge, and average General Language Understanding (GLU), measured in percentage points. Balanced Mixture (BM) uses the 13 languages shared by FINEWEB2-HQ-QA and TDQA with equal token allocation and Natural Mixture (NM) follows each dataset's natural language distribution. *Ours* denotes FINEWEB2-HQ-QA, *(T)DQA* denotes DQA+TDQA, and *(T)DQA+Ours* adds FINEWEB2-HQ-QA. Each configuration is trained once. Error bars show ± 1 standard error of the paired difference from eval sampling only, assuming independence across benchmarks and clustering translations of the same item across languages within translated benchmarks. Comparisons should be made within panels, since baselines and evaluation language sets differ.

prior findings that pretraining data augmentation can improve knowledge extraction (Allen-Zhu & Li, 2024).

For General Language Understanding (GLU), gains are positive in three of the four settings, with a small decrease of 0.2 points in 3B BM. At 3B, native QA approximately preserves GLU in BM and improves it in NM. At 8B, native QA shows positive GLU trends in both settings, with gains of 0.9 points in BM and 0.5 in NM. However, per-benchmark GLU metrics reveal a more nuanced picture. In the Balanced Mixture setting, native QA decreases XWinograd at both model scales, and PAWS-X also drops at 3B (Table 5). These per-benchmark differences show that positive category-level trends can coexist with lower scores on individual tasks. Thus, the aggregate GLU results do not imply uniform improvements across its constituent tasks.

Takeaway: FINEWEB2-HQ-QA improves both knowledge categories over the non-QA baseline across all four settings, while largely preserving or improving aggregate GLU.

5.3 PERFORMANCE PROFILES OF PUBLIC QA RESOURCES UNDER MATCHED BUDGETS (Q2)

To answer this question, we focus on the language-controlled *Balanced Mixture* (BM) setup in Figure 1 and Tables 5 and 6. This setup reduces language-mixture differences by distributing the synthetic budget equally over the 13 shared target languages for each multilingual QA source. DQA is English-only by construction and is therefore used as an English-source QA control.

At the 3B scale, English QA (DQA) and translated QA (TDQA) show higher observed Global Knowledge scores than native QA. TDQA provides the largest observed Global Knowledge gain among the single-source QA datasets, improving over the baseline by 12.8 points. This likely reflects, in part, the knowledge-focused construction of its English-source QA data, generated from high-quality web documents. Su et al. (2025) show that the high-quality Nemotron subset used as the basis for DQA improves MMLU by 5.6 points over DCLM (Li et al., 2024) for an 8B model trained on 1T tokens. Native QA exhibits a different profile. Although its Global Knowledge gains are smaller than those of DQA and TDQA at 3B, it shows the most favorable GLU behavior among

the QA sources, largely preserving the baseline macro-average and reducing the GLU degradation observed with English-derived QA. One possible explanation is that FINEWEB2-HQ-QA is closer to the language and domain distribution of the baseline mixture, as it is generated from native-language documents rather than knowledge-dense English-source QA. This may reduce distribution shift for general language understanding, while also making the data less concentrated on the broad academic and professional knowledge targeted by benchmarks such as MMLU.

At 8B, native QA becomes more competitive, especially on Local Knowledge and GLU. TDQA shows the largest observed Global Knowledge gain (+4.5 over baseline), while native QA shows the largest observed Local Knowledge gain (+3.6). Native QA also remains among the strongest on observed GLU Avg. One possible explanation for this setup dependence is that English-derived QA provides a concentrated global factual signal that is particularly useful to smaller, less extensively pretrained models, whereas greater capacity and more extensive pretraining may help models exploit the complementary knowledge available in target-language document-grounded QA.

Combining QA sources yields differences between training regimes. At 3B, the union of English, translated, and native QA achieves scores close to or above those of the best individual source across evaluation categories, suggesting that mixing sources can maintain broad performance under a fixed token budget at smaller scale. At 8B, the combination largely preserves the strongest GLU behavior but underperforms the best single source on both knowledge categories. Since the fixed 30B budget is split equally among three sources, the strongest contributor per metric (TDQA for Global Knowledge, native QA for Local Knowledge) receives only a third of the tokens it would in isolation. This reduced exposure may contribute to the observed pattern, although our experiments do not isolate the effect of per-source token allocation. Thus, in this setup, combining native and English-derived QA yields different performance trade-offs, with the benefits of mixing varying across training regimes and budget allocations.

For completeness, Appendix E.2 reports category-specific per-language breakdowns for the 8B BM model. These results are consistent with the aggregate trends, while also showing language-level variation.

> **Takeaway:** Nemotron's Translated QA shows the highest observed Global Knowledge scores among single sources, while native QA shows a different profile, with competitive GLU and strong Local gains at 8B. These differences may reflect both the generation strategy and the knowledge content and density of the underlying source corpora.

5.4 DATASET USAGE IN PRACTICE (Q3)

The *Natural Mixture* setting is closest to practical multilingual pre-training, since it preserves each dataset's natural language distribution and uses all available languages. In this setting, FINEWEB2-HQ-QA is highly effective on its own: it improves both knowledge categories at 3B and 8B while showing the highest observed GLU Avg among single-source QA datasets.

In both 3B and 8B settings, TDQA shows higher observed Global Knowledge scores than native QA but lower aggregate GLU. Combining sources yields a more balanced observed profile across knowledge and GLU. At 3B, the three-way combination achieves the largest observed gains in both knowledge categories, although these are close to those of (T)DQA within evaluation uncertainty, while also improving GLU over the baseline. At 8B, the combination does not exceed the strongest individual source in either knowledge category, but retains Global Knowledge gains between those of TDQA and native QA while maintaining GLU close to native QA. Taken together, these results suggest that the combined mixture can retain strong knowledge performance alongside favorable aggregate GLU, but does not consistently outperform the strongest individual source.

Appendix E.3 reports CulturalBench results. Although the benchmark is volatile due to small per-country sample sizes, the combined mixture shows the highest observed aggregate scores in both mixture settings, a pattern consistent with the different strength profiles of the individual sources. With few items per country and single-seed training we treat CulturalBench as exploratory only.

> **Takeaway:** Under natural-mixture settings, native QA improves both knowledge categories and aggregate GLU over the non-QA baseline. Combining native, English, and translated QA

changes the balance across these categories, but does not consistently outperform the strongest individual source.

6 Limitations and Conclusion

We introduced FineWeb2-HQ-QA, a 193B-token native document-grounded multilingual QA dataset covering 20 languages, with positive knowledge gains in all four evaluated settings and generally favorable GLU results. As a standalone augmentation, it shows strong observed general language understanding and Local Knowledge in our settings, particularly at 8B. Relative to native QA, Nemotron's translated QA has higher observed Global Knowledge and lower benchmark-averaged GLU scores in all four settings. These descriptive differences may reflect source content or generation-pipeline differences. Under natural-mixture settings, combining native and Nemotron QA yields a balanced observed profile across knowledge and general language understanding, without consistently exceeding the strongest alternative in each category. Overall, native document-grounded QA provides a complementary profile to English-derived QA, whether used in English or translated, and presents a valuable addition to publicly released resources for multilingual pre-training.

Our results also highlight directions for future work. Controlled comparisons could clarify how source quality, topical distribution, and knowledge density affect the utility of synthetic QA. Manual checks revealed logical errors, suggesting that a verification step (Liu et al., 2025) would be a promising addition.

More generally, native QA requires high-quality open-source models capable of generating linguistically rich and reliable data from multilingual inputs.

Our findings suggest that substantial progress is still possible in closing the gap with English-language resources and improving accessibility across diverse cultures and languages. To support further research, we are releasing our dataset and codebase.

AI use statement

AI models were used for synthetic QA generation, the automated quality audit, and English translations of selected examples, as described in the paper. AI was also used to improve manuscript grammar and clarity.

References

Loubna Ben Allal, Anton Lozhkov, and Daniel van Strien. Cosmopedia: how to create large-scale synthetic data for pre-training. *Hugging Face Blog*, pp. 56, 2024.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of Language Models: Part 3.1, Knowledge Storage and Extraction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML '24, July 2024. Full version available at https://ssrn.com/abstract=5250633.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, et al. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*, 2025.

Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, et al. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, et al. Nvidia nemotron 3: Efficient and open intelligence. *arXiv preprint arXiv:2512.20856*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Aldo Gael Carranza, Kaleigh Mentzer, Ricardo Pio Monti, Alex Fang, Alvin Deng, Amro Abbas, Anshuman Suri, Brett Larsen, Cody Blakeney, Darren Teh, et al. \" uberweb: Insights from multilingual curation for a 20-trillion-token dataset. *arXiv preprint arXiv:2602.15210*, 2026.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms, 2024. URL https://arxiv.org/abs/2410.02677.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pp. arXiv–2307, 2023.

Ona De Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer Van Der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, et al. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 1116–1128, 2024.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 1286–1305, 2021.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations, 2024. *URL https://arxiv. org/abs/2406.08446*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna B Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10750–10769, 2025.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6155–6218, 2023.

Sakaguchi Keisuke, Le Bras Ronan, Bhagavatula Chandra, and Choi Yejin. Winogrande: An adversarial winograd schema challenge at scale. 2019.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions. *Transactions of the Association for Computational Linguistics*, 13:1032–1055, 2025.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you! 2023.

Zhengzhong Liu, Liping Tang, Linghao Jin, Haonan Li, Nikhil Ranjan, Desai Fan, Shaurya Rohatgi, Richard Fan, Omkar Pangarkar, Huijuan Wang, et al. K2-v2: A 360-open, reasoning-enhanced llm. *arXiv preprint arXiv:2512.06201*, 2025.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. Enhancing multilingual llm pretraining with model-based data selection. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pp. 31–56, 2025.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2022.

MultiSynt. Mt-nemotron-cc: Large-scale machine-translated high quality web text, 2025. URL https://huggingface.co/datasets/MultiSynt/MT-Nemotron-CC. A translated variant of Nemotron-CC High Quality for multilingual LLM pretraining.

Raymond Ng, Thanh Ngan Nguyen, Huang Yuli, Tai Ngee Chia, Leong Wai Yi, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, et al. Sea-lion: Southeast asian languages in one network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 512–526, 2025.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, et al. Seallms-large language models for southeast asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 294–304, 2024.

Joel Niklaus, Guilherme Penedo, Hynek Kydlicek, Elie Bakouch, Lewis Tunstall, Ed Beeching, Thibaud Frere, Colin Raffel, Leandro von Werra, and Thomas Wolf. The synthetic data playbook: Generating trillions of the finest tokens, 2026.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Matteo Pagliardini, Pierre Ablin, and David Grangier. The ademamix optimizer: Better, faster, older. *arXiv preprint arXiv:2409.03137*, 2024.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset, Aug 2016.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *arXiv preprint arXiv:2506.20920*, 2025.

PleIAs. Synth: The new data frontier, November 2025. URL https://pleias.fr/blog/blogsynth-the-new-data-frontier. Accessed: 2026-03-18.

Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://ducdauge.github.io/files/xcopa.pdf.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2459–2475, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Alexey Tikhonov and Max Ryabinin. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning, 2021.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: Main volume*, pp. 2203–2213, 2021.

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

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL https://aclanthology.org/D19-1382.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P. Xing. Megamath: Pushing the limits of open math corpora. *arXiv preprint arXiv:2504.02807*, 2025. Preprint.

A LIMITATIONS

Experimental scope. Each configuration was trained once and reported standard errors quantify evaluation-set uncertainty rather than run-to-run training variance. Aggregate results need not hold uniformly for every task or language.

Dataset coverage. FINEWEB2-HQ-QA covers 20 languages and draws its content from selected FineWeb2-HQ documents. Here, *native* refers to target-language sourcing and generation, not necessarily original authorship in that language. The resulting QA coverage is shaped by the source corpus and document selection and does not imply representation of all communities using each language.

Residual data-quality issues. Our filtering pipeline addresses selected output-quality issues but does not verify the correctness of every generated QA pair. The automated audit in Appendix F provides initial evidence on data quality. The audit's repeatability and injected-error tests support consistency and sensitivity to the tested defects. These assessments are limited in coverage. A broader, more in-depth evaluation across languages and error types remains a direction for future work.

B TRAINING SETUP

As outlined in the main text, our cooldown experiments utilize a 100B-token training pool constructed from 30B tokens of the target dataset and 70B tokens from a baseline mixture. The exact composition and sampling weights of this baseline mixture are detailed in Table 2.

Dataset	Proportion	Reference
DCLM-EDU	53%	(Allal et al., 2025)
FineWeb2-HQ (top 10%)*	33%	(Apertus et al., 2025)
Starcoder	10%	(Li et al., 2023)
MegaMath-Web	4%	(Zhou et al., 2025)

Table 2: 70B-token base mixture, adapted from Apertus (Apertus et al., 2025) Phase 4, with FineWeb2-HQ centered on high-resource European languages (denoted with *).

We run all experiments using `Megatron-LM` (Shoeybi et al., 2019). We keep training hyperparameters consistent across experiments within each configuration.

8B model. We follow Apertus et al. (2025) and continue training from their 9.5T token checkpoint. Starting from this checkpoint, we begin an immediate cooldown phase over 63B tokens using a 1-sqrt decay schedule (Hägele et al., 2024), annealing the learning rate to $0.1\times$ its peak value. We keep the peak learning rate at 1.1×10^{-4} and retain AdEMAMix (Pagliardini et al., 2024) as the optimizer.

3B model. Starting from the 200B-token base checkpoint described in Section C, we cool down for 50B tokens. Relative to the 8B configuration, we adopt an analogous training configuration with the same general recipe but adjust several hyperparameters to suit the smaller model scale. Specifically, we reduce the global batch size from $\sim$4M to $\sim$2M tokens per batch while retaining a sequence length of 4096, and increase the peak learning rate to 1.5×10^{-4}. We keep the warmup-stable-decay (WSD) schedule (Hu et al., 2024; Hägele et al., 2024), cooling down to a minimum learning rate of 0. The architecture consists of 28 layers with a hidden size of 3072, an FFN intermediate size of 12288, 24 attention heads, and a RoPE scaling factor of 32.

C SUPPORTING NEMOTRON SUBSET ABLATIONS

The release of NVIDIA Nemotron 3 (Blakeman et al., 2025) demonstrated highly competitive performance, partially attributed to its high-quality pre-training data. Consequently, its publicly released data, Nemotron-CC-v2 (Basant et al., 2025), provides a natural starting point for investigating

effective ways to augment the FineWeb2-HQ dataset. We ablate the subsets of Nemotron-CC-v2 (Basant et al., 2025) to inform our synthetic refinement strategy. These ablations follow the training and evaluation setup described in Section 4.

3B Cooldown Results. As detailed in Table 3, QA datasets have a clear impact on English performance, notably more pronounced than including the highest-quality subset or its rephrased counterpart. Because our focus is on augmenting natively curated datasets effectively within an allocated token budget, QA augmentation emerges as a highly efficient strategy. Interestingly, QA-augmented documents boost performance independently of the target language. However, small models trained on as few as 250B tokens often lack the capacity to reliably answer multiple fixed-choice (MFC) questions (Gu et al.). Thus, these gains may partly reflect improved MFC answering rather than broader knowledge acquisition. We therefore validate this finding at the 8B scale.

Model	English Avg	Multi. Avg
Baseline	48.6 ±0.4	45.5 ±0.4
High Quality	49.7 ±0.4	45.1 ±0.4
High Quality Synthetic	49.0 ±0.4	45.0 ±0.4
Diverse QA	**54.7** ±0.4	47.2 ±0.4
Translated Diverse QA	51.4 ±0.4	**47.9** ±0.4

Table 3: 3B cooldown performance across base, high-quality, and synthetic mixtures. Scores are macro-averaged accuracy (%) over English and multilingual benchmarks ± one standard error from eval sampling only, assuming independence across benchmarks and clustering translations of the same item across languages within translated benchmarks. Each configuration is trained once.

8B Cooldown Results. As shown in Table 4, testing these subsets at the 8B scale confirms that QA-augmented documents continue to improve both English and Multilingual performance overall. Note that for compute cost reasons, we omit the High Quality and High Quality Synthetic ablations at this scale, focusing on the QA subsets.

Interestingly, Nemotron's English QA subset performs comparably to its translated counterpart on the multilingual macro-average, suggesting that English QA augmentation can also improve multilingual evaluation scores. This motivates exploring what benefits QA generated from native-language source documents, which reflect different topical distributions, can offer.

Model	English Avg	Multi. Avg
Baseline	62.0 ±0.4	55.1 ±0.3
DQA	**64.8** ±0.4	**56.0** ±0.3
TDQA	63.4 ±0.4	55.9 ±0.3

Table 4: 8B cooldown performance across the baseline and QA-augmented mixtures. Scores are macro-averaged accuracy (%) over English and multilingual benchmarks ± one standard error from eval sampling only, assuming independence across benchmarks and clustering translations of the same item across languages within translated benchmarks. Each configuration is trained once.

D Dataset Generation

D.1 QA Augmentation

Below we show the prompt template used to generate our native QAs.

Generation Prompt Template

Task: Read the text, ask questions and answer them.

LANGUAGE CONSTRAINTS:

1. The questions and answers themselves must be written in {language}.

2. The labels "Question:" and "Answer:" must remain in English.

Follow these instructions:

1. Ask diverse questions that require different cognitive skills or cover different aspects of the text.
2. Ask questions in various forms such as:
   - Yes/No questions that require determining whether a statement is true or false.
   - Open-ended questions that begin with words like what, how, when, where, why and who in target language {language}.
   - Multi-choice questions that offers two or more options to choose from. Include the options in the question.
   - Comparison questions that compare two quantities or objects and determine the relationship between them.
   - Reading comprehension questions that test the ability to understand and analyze the text.
   - Problem-solving questions that test the ability to solve mathematical, physical, or logical problems.
3. Focus on asking questions about factual information, important knowledge, or concrete details in the text.
4. Write questions and answers using clear and concise language.
5. Use plain text. Do not use Markdown.
6. Each question and answer pair should be on a separate line. Tag the question in {language} with "Question:" and the answer in {language} with "Answer:".

Text:
{text_placeholder}

Task:
After reading the above {language} text, ask up to 8 questions and provide the correct answers following the instructions.

Give your response in this format:
Here are the questions and answers based on the provided text:
- Question: [first question in {language}] Answer: [first answer in {language}]
- Question: [second question in {language}] Answer: [second answer in {language}]

D.2 Filtering Details

Below we detail each of the four filtering steps. Overall and per-language dataset statistics are reported in Table 1.

Aborted Reasoning Traces (ART). We filter all samples, where the `</think>` closing tag is missing, assuming that a completed reasoning trace always contains it. This step removes 1.87% of the data, as shown in Table 1 alongside the per-language filtering ratios.

Language Mismatch (LID). To catch cases where the model generates in the wrong language (e.g., defaulting to English), we enforce a minimum language identification score of the QA pairs using GlotLID (Kargaran et al., 2023), the same LID system utilized for FineWeb2 (Penedo et al., 2025). For simplicity, we apply a uniform threshold across languages. To calibrate this threshold, we plot the LID score distribution and select *0.9* as cutoff (Appendix D.3; Figure 2). The generated QA pairs are largely consistent with the requested language. This step removes 1.24% of the data (see Appendix D.4 for a removed Arabic sample).

Degenerate Samples (DS). High local repetition has been identified as a factor that degrades pre-training quality and is hypothesized to cause LLMs to repeat themselves during sampling (Rae et al., 2021). Therefore, we apply the Gopher repetition filter, targeting samples where the same question is asked multiple times or phrases are repeated verbatim. Similar to language mismatch filtering, we inspect the per-language distribution of filter scores (Appendix D.3; Figure 3) and select

a threshold validated by manual sanity checks. Notably, the Japanese distribution exhibits high repetition scores, which may reflect issues with generation quality. Overall, this step removes 4.64% of the data (see Appendices D.4 and D.4 for Italian and Japanese examples).

Benchmark Contamination (BC). Since we generate QA pairs synthetically, there is a risk that the model generates benchmark QAs rather than QAs relevant to the sample text. While prompting directly on the native FineWeb2-HQ documents helps mitigate this behavior, we run a decontamination pipeline against the key benchmarks used in our evaluation. We employ the standard 13-gram hash decontamination pipeline from `datatrove`, adapted from `lighteval` to fit our `lm-harness` evaluation setup. This step removes 0.09% of the data (see Appendix D.4 for an Indonesian sample).

D.3 Filter Score Distributions

Figure 2 shows the language identification (LID) score distribution from GlotLID (Kargaran et al., 2023) for 1K random samples per-language. This distribution was used to select the threshold of 0.9 for the *Language Mismatch* filtering step. Figure 3 shows the Gopher n-gram and paragraph/line repetition score distributions for up to 4K random samples per-language. We aggregate all n-gram and paragraph/line scores by taking the maximum within each group, i.e. the most sensitive score and the one most likely to trigger filtering. We use this overview to select the 0.25 and 0.1 thresholds for the n-gram and paragraph/line *Degenerate Samples* filtering steps, respectively.

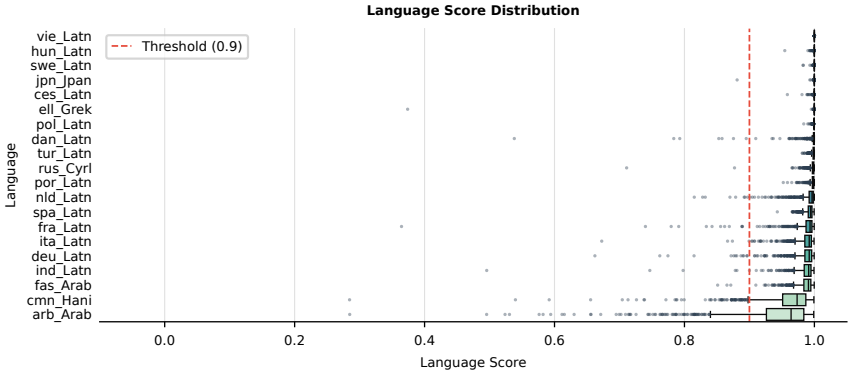


Figure 2: Language identification (LID) score distribution from GlotLID (Kargaran et al., 2023) over 1K randomly selected documents from FineWeb2-HQ (Messmer et al., 2025) after the *Aborted Reasoning Traces* filtering step. Higher scores indicate greater confidence that the document matches its assigned language label.

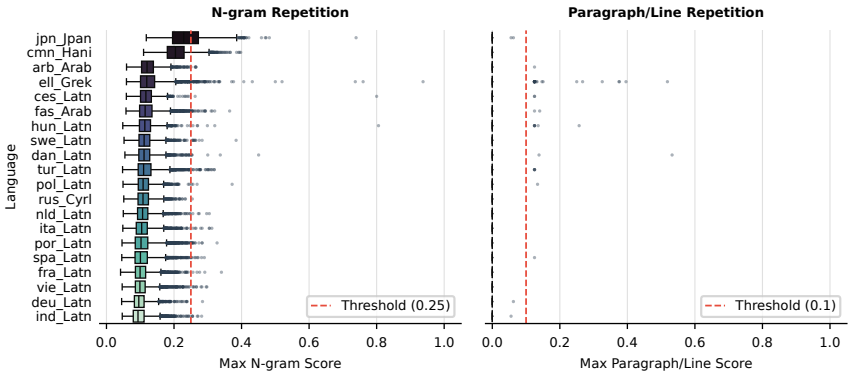


Figure 3: Gopher n-gram and paragraph/line repetition score distributions up to 4K randomly selected documents from FineWeb2-HQ (Messmer et al., 2025) after the *Aborted Reasoning Traces* and *Language Mismatch* filtering steps. Each score is the maximum within its group (n-gram or paragraph/line), representing the most sensitive filter for that sample.

D.4 Removed Samples

In this section, we discuss randomly selected samples removed by the *Language Mismatch*, *Degenerate Samples*, and *Benchmark Contamination* pipeline steps, focusing on languages with out-of-distribution scores as observed in Figures 2 and 3 in Section D.3.

Language Recognition Figure 2 shows that Arabic is the language with the overall lowest language identification scores. We asked Gemini 3 Pro to translate the samples to English as closely as possible to the originals. While there appears to be an error in pronoun usage, the samples generally look good. These inspected examples suggest that the threshold may also exclude valid Arabic samples. Although this step removes approximately 1% of samples overall, it removes 16.96% of Arabic samples. We retain the uniform threshold, while acknowledging its disproportionate impact on Arabic and the possibility of false exclusions.

Original (Arabic)

- Question: هل الطبقة الأولى مكونة من الكيك؟ Answer: نعم
- Question: ما هي الخطوة الأولى في تحضير الطبقة الأولى؟ Answer: خذ مقدار من الكيك ورصيه في قعر كاسات التقديم
- Question: ما المكونات المطلوبة للطبقة الثانية؟ Answer: جبن الكريم والقشدة
- Question: ما الغرض من الخطوة الرابعة؟ Answer: تجميد المزيج في الثلاجة
- Question: ما الفرق بين الطبقة الثالثة والطبقة الثانية؟ Answer: الطبقة الثالثة تحتوي على الشوكولاتة بينما الطبقة الثانية تحتوي على جبن الكريم
- Question: كم دقيقة يجب تجميد الطبقة الثالثة؟ Answer: 30 دقيقة
- Question: ما المكون الذي يُضاف إلى الشوكولاتة في الطبقة الثالثة؟ Answer: القشدة
- Question: ما الخطوة التي تلي تحضير الطبقة الثانية؟ Answer: تجميد الكاسات في الثلاجة

Translation (English)

```
- Question: Is the first layer made of cake? Answer: Yes.
- Question: What is the first step in preparing the first layer?
Answer: Take [sir] amount of cake and arrange [madam] it in the
bottom of the serving cups.
- Question: What are the required ingredients for the second layer?
Answer: Cheese of cream and table cream.
- Question: What is the purpose of the fourth step? Answer:
Freezing the mixture in the fridge.
- Question: What is the difference between the third layer and the
second layer? Answer: The third layer contains chocolate while the
second layer contains cheese of cream.
- Question: How many minutes must the third layer be frozen? Answer:
30 minutes.
- Question: What is the ingredient added to the chocolate in the
third layer? Answer: Table cream.
- Question: What is the step that follows preparing the second layer?
Answer: Freezing the cups in the fridge.
```

Line repetition Figure 3 shows that line and paragraph repetition occurs rarely across all languages. For language diversity, we select Italian as a representative example. We observe that Questions 2 and 8 are indeed identical replicas, representing a valid exclusion.

Original (Italian)

```
- Question: Messi ha vinto il Pallone d'Oro 2019? Answer: Sì.
- Question: Chi ha vinto il Pallone d'Oro 2019? Answer: Lionel
```

Messi.
- Question: Chi è arrivato quarto nella classifica del Pallone d'Oro 2019? Answer: Sadio Mané.
- Question: Chi ha vinto il Pallone d'Oro 2019, Messi o Virgil Van Dijk? Answer: Messi.
- Question: Perché Messi ha espresso dispiacere per Sadio Mané? Answer: Perché Mané ha finito al quarto posto.
- Question: Perché Messi ha votato per Sadio Mané? Answer: Per il suo modo di giocare e le prestazioni con il Liverpool e la Nazionale.
- Question: Messi ha detto che Mané era il migliore? Answer: No, ha detto che c'erano tanti giocatori meritevoli.
- Question: Chi ha vinto il Pallone d'Oro 2019? Answer: Lionel Messi

Translation (English)
- Question: Did Messi win the 2019 Ballon d'Or? Answer: Yes.
- Question: Who won the 2019 Ballon d'Or? Answer: Lionel Messi.
- Question: Who finished fourth in the 2019 Ballon d'Or rankings? Answer: Sadio Mané.
- Question: Who won the 2019 Ballon d'Or, Messi or Virgil van Dijk? Answer: Messi.
- Question: Why did Messi express regret for Sadio Mané? Answer: Because Mané finished in fourth place.
- Question: Why did Messi vote for Sadio Mané? Answer: For his style of play and his performances with Liverpool and the national team.
- Question: Did Messi say that Mané was the best? Answer: No, he said there were many deserving players.
- Question: Who won the 2019 Ballon d'Or? Answer: Lionel Messi.

5-gram duplication Figure 3 shows high n-gram repetition for Japanese. To inspect whether this is a valid removal, we again asked Gemini 3 Pro to translate the sample as closely as possible while retaining errors. We observe a distinct question-answer pattern in which the answer nearly verbatim repeats the question. We consider this a low-quality generation. The high n-gram repetition ratio compared to other languages and low performance in the per-language aggregation in Table 11 of our native QA suggest that the Japanese QAs may suffer in quality.

Original (Japanese)
- Question: このサイトは独学向けの英語学習サイトですか？Answer: はい、このサイトは独学でも身につく英語学習サイトです。
- Question: 受動態の形式はどのようなものですか？Answer: 受動態の形式は「be動詞＋過去分詞」です。
- Question: このページの読了時間はどれくらいですか？Answer: このページの読了時間は約0分0秒です。
- Question: 他の関連記事のテーマは何ですか？Answer: 他の関連記事のテーマは受動態、shouldの用法、現在分詞です。
- Question: 現在分詞の形はどのように表されますか？Answer: 現在分詞の形は「V-ing」です。
- Question: このサイトの主な目的は何ですか？Answer: このサイトの主な目的は英語学習をサポートすることです。
- Question: 読了時間が0分0秒であることは正しいですか？Answer: はい、このページの読了時間は約0分0秒と記載されています。
- Question: 分詞の定義は何ですか？Answer: 分詞とは「動詞と形容詞の性質を合わせ持った複合語」のことです。

Translation (English)

```
- Question:  Is this an English learning website designed for
self-study?  Answer:  Yes, this is an English learning website where
you can acquire skills even through self-study.
- Question:  What is the form of the passive voice?Answer:  The form
of the passive voice is "be verb + past participle".
- Question:  What is the estimated reading time for this page?Answer:
The reading time for this page is approximately 0 minutes and 0
seconds.
- Question:  What are the topics of the other related articles?Answer:
The topics of the other related articles are the passive voice, the
usage of "should," and present participles.
- Question:  How is the present participle form expressed?Answer:  The
present participle form is expressed as "V-ing".
- Question:  What is the main purpose of this website?Answer:  The
main purpose of this website is to support English learning.
- Question:  Is it correct that the reading time is 0 minutes and 0
seconds?Answer:  Yes, the reading time for this page is stated to be
approximately 0 minutes and 0 seconds.
- Question:  What is the definition of a participle?Answer:  A
participle is defined as a "compound word that combines the properties
of both a verb and an adjective.
```

Decontamination Benchmark decontamination removes below 1% of samples across all languages. For language diversity, we show an Indonesian example. The contaminated sequence “a 0 dan 0 b 0 dan 0 c 0 dan 0 d” from the INCLUDE benchmark appears in QAs about an image, which suggests a possible false-positive match.

Original (Indonesian)

```
Batang baja P - Q dibuat menjadi magnet dan setelah menjadi magnet,
beberapa kompas diletakkan di sekitar batang baja seperti gambar.
Posisi kompas yang benar ditunjukkan oleh gambar ....
A. 1 dan 2
B. 1 dan 3
C. 2 dan 4
D. 3 dan 4
Jawab:  A
Pertama perlu mengetahui kutub magnet pada ujung batang baja P - Q
terlebih dahulu.  Cara menentukan kutub magnet pada ujung batang baja
P - Q dapat menggunakan kaidah tangan kanan.
Arus listrik mengalir dari muatan positif ke muatan negatif.  Gambaran
arus listrik yang mengalir pada batang baja P - Q sesuai dengan
kondisi berikut.
Selanjutnya, sobat idschool dapat menggunakan kaidah tangan kanan
untuk mengetahui mana kutub utara mana kutub selatan.  Di mana empat
jari (selain ibu jari) menunjuk arah arus listrik dan ibu jari selalu
menunjuk kutub utara.
Akan diperoleh kutub utara batang baja P - Q berada di titik P dan
kutub selatan di titik Q. Medan magnet di sekitar baja kutub utara
ke kutub selatan sehingga arah medan magnet untuk batang baja P - Q
sesuai gambar di bawah.
```

Kutub kompas yang mendekati batang baja P – Q akan berlawanan dengan muatan ujung baja P – Q. Selanjutnya perlu mengetahui posisi kompas yang benar ditunjukkan oleh gambar yang sesuai.
Posisi kompas yang benar ditunjukkan oleh gambar dapat dianalisa dengan cara berikut.
- Kompas 1:
Kutub kompas yang berdekatan dengan titik P adalah kutub selatan (warna putih) → Benar
- kompas 2:
Kutub kompas yang berdekatan dengan titik P adalah kutub selatan (warna putih) → Benar
- Kompas 3:
Medan magnet yang mengenai ujung kompas nomor 3 dari kiri adalah kutub utara sehingga ujung kompas nomor tiga bagian kiri adalah selatan (warna putih) → Salah, karena gambar ujung kompas nomor 3 bagian kiri berwarna hitam (seharunya berwarna putih).
- Kompas 4:
Titik Q merupakan kutub selatan sehingga ujung bagian kompas yang berdekatan dengan titik Q adalah kutub utara (warna hitam) → Salah, karena pada gambar kompas nomor 4 memiliki ujung yang berdekatan dengan titik Q berwarna hitam
Jadi, posisi kompas yang benar ditunjukkan oleh gambar 1 dan 2.

- Question: Apakah posisi kompas yang benar ditunjukkan oleh gambar 1 dan 2? Answer: Ya
- Question: Bagaimana cara menentukan kutub utara dan selatan batang baja P – Q? Answer: Dengan menggunakan kaidah tangan kanan.
- Question: Mengapa kompas 3 dianggap salah dalam teks? Answer: Karena ujung kompas 3 bagian kiri seharusnya berwarna putih (kutub selatan) tetapi dalam gambar berwarna hitam. ...

Translation (English)

The steel bar P – Q is magnetized, and after becoming a magnet, several compasses are placed around the steel bar as shown in the picture.
The correct compass positions are shown by pictures .... A. 1 and 2 B. 1 and 3 C. 2 and 4 D. 3 and 4
Answer: A
First, we need to determine the magnetic poles at the ends of the steel bar P – Q. The right-hand rule can be used to determine the magnetic poles at the ends of the steel bar P – Q.
Electric current flows from positive charge to negative charge. The illustration of the electric current flowing in the steel bar P – Q corresponds to the following conditions.
Next, you can use the right-hand rule to find out which is the north pole and which is the south pole. The four fingers (other than the thumb) point in the direction of the electric current, and the thumb always points to the north pole.
You will find that the north pole of the steel bar P – Q is at point P and the south pole is at point Q. The magnetic field around the steel goes from the north pole to the south pole, so the direction of the

```
magnetic field for the steel bar P - Q is according to the picture
below.
The compass pole approaching the steel bar P - Q will be opposite
to the polarity of the end of the steel bar P - Q. Next, we need to
identify the correct compass positions shown by the corresponding
pictures.
The correct compass positions shown by the pictures can be analyzed in
the following way:
Compass 1:  The compass pole adjacent to point P is the south pole
(white color) → Correct
Compass 2:  The compass pole adjacent to point P is the south pole
(white color) → Correct
Compass 3:  The magnetic field hitting the left end of compass number
3 is the north pole, so the left end of compass number three is the
south pole (white color) → Incorrect, because the picture of the left
end of compass number 3 is black (it should be white).
Compass 4:  Point Q is the south pole, so the part of the compass
end adjacent to point Q is the north pole (black color) → Incorrect,
because in the picture, compass number 4 has the end adjacent to point
Q colored black.
So, the correct compass positions are shown by pictures 1 and 2.
- Question:  Are the correct compass positions shown by pictures 1 and
2?  Answer:  Yes
- Question:  How do you determine the north and south poles of the
steel bar P - Q? Answer:  By using the right-hand rule.
- Question:  Why is compass 3 considered incorrect in the text?
Answer:  Because the left end of compass 3 should be white (south
pole), but in the picture, it is black.  ...
```

E Additional Results

E.1 Benchmark-Level Aggregate Results

Tables 5 to 8 report benchmark-level aggregate results for each evaluation category at 3B and 8B scale. Tables 5 and 6 show the *Balanced Mixture (BM)* setting, while Tables 7 and 8 show the *Natural Mixture* setting.

	Global Knowledge	*Local Knowledge*	*General Language Understanding*				
Model	**global_mmlu**	**include_base_44**	**Avg**	**xnli**	**pawsx**	**xcopa**	**xwinograd**
Baseline (BM)	25.9	25.4	58.2	43.1	55.4	65.7	68.8
DQA	34.3	33.7	57.4	43.4	55.5	64.0	66.7
TDQA (BM)	38.7	33.4	57.6	42.7	54.3	65.3	68.1
Ours (BM)	30.8	27.8	58.0	43.6	54.8	65.6	67.9
(T)DQA (BM)	39.4	34.9	57.5	43.2	54.1	65.8	66.8
(T)DQA + Ours (BM)	39.3	34.5	57.9	43.1	54.2	66.0	68.2

Table 5: Performance of 3B models using *Balanced Mixture (BM)* by benchmark. Each benchmark is averaged over its covered languages. Avg further averages these benchmark scores. Each configuration trained once. All metrics are accuracy.

	Global Knowledge	*Local Knowledge*	*General Language Understanding*						
Model	**global_mmlu**	**include_base_44**	**Avg**	**xnli**	**pawsx**	**m_arc***	**m_hellaswag***	**xcopa**	**xwinograd**
Baseline (BM)	54.8	51.1	61.6	44.4	59.4	46.2	68.3	75.7	75.9
DQA	56.6	53.8	62.4	44.0	60.5	49.8	68.1	75.7	76.1
TDQA (BM)	59.3	53.4	61.6	43.2	59.5	50.3	67.9	74.5	74.2
Ours (BM)	57.5	54.7	62.5	44.7	61.3	50.2	68.3	75.9	74.8
(T)DQA (BM)	58.5	53.8	61.8	43.6	59.7	50.4	67.7	74.5	75.1
(T)DQA + Ours (BM)	57.0	53.5	62.3	44.0	60.6	50.5	68.3	75.5	75.0

Table 6: Performance of 8B models using *Balanced Mixture (BM)* by benchmark. Each benchmark is averaged over its covered languages. Avg further averages these benchmark scores. Each configuration trained once. All metrics are accuracy, with (*) denoting normalized accuracy.

	Global Knowledge	*Local Knowledge*	*General Language Understanding*				
Model	**global_mmlu**	**include_base_44**	**Avg**	**xnli**	**pawsx**	**xcopa**	**xwinograd**
Baseline	24.8	25.0	55.8	41.5	56.8	55.4	69.4
DQA	32.8	30.8	55.0	40.9	55.5	55.1	68.4
TDQA	35.2	29.5	55.6	41.5	56.0	56.1	69.0
Ours	30.4	26.5	57.1	42.5	56.5	57.8	71.7
(T)DQA	36.8	31.1	54.9	41.1	55.6	56.0	66.9
(T)DQA + Ours	37.2	31.4	56.4	42.5	55.4	57.3	70.4

Table 7: Performance of 3B models using *Natural Mixture* by benchmark. Each benchmark is averaged over its covered languages. Avg further averages these benchmark scores. Each configuration trained once. All metrics are accuracy.

	Global Knowledge	*Local Knowledge*	*General Language Understanding*						
Model	**global_mmlu**	**include_base_44**	**Avg**	**xnli**	**pawsx**	**m_arc***	**m_hellaswag***	**xcopa**	**xwinograd**
Baseline	50.0	47.0	57.3	44.4	61.0	38.9	54.8	65.5	79.2
DQA	52.0	50.6	57.6	44.3	61.0	41.4	54.7	65.4	78.9
TDQA	54.2	50.1	57.2	43.7	60.5	41.1	54.9	64.9	78.1
Ours	52.6	50.4	57.8	44.8	61.3	41.0	54.8	65.5	79.2
(T)DQA	52.4	50.6	57.2	43.5	60.5	41.6	54.8	65.1	77.9
(T)DQA + Ours	53.6	50.1	57.7	44.1	60.6	42.0	54.9	65.4	79.1

Table 8: Performance of 8B models using *Natural Mixture* by benchmark. Each benchmark is averaged over its covered languages. Avg further averages these benchmark scores. Each configuration trained once. All metrics are accuracy, with (*) denoting normalized accuracy.

E.2 PER-LANGUAGE AGGREGATE RESULTS

Tables 9 to 11 report per-language category aggregates for the 8B BM setting. N/A denotes unavailable benchmark coverage, and macro averages exclude unavailable entries. For GLU, each language score averages over the benchmarks available for that language, and Macro Avg then averages across languages. In contrast, the main GLU Avg in Table 6 averages over languages within each benchmark before averaging across benchmarks. Because benchmark language coverage varies, these aggregation procedures assign different weights to the available scores and can yield different averages and rankings.

Model	Arabic	Chinese	Danish	German	French	Italian	Japanese	Dutch	Polish	Portuguese	Russian	Spanish	Swedish	Macro Avg
Baseline (BM)	49.0	52.5	N/A	57.0	57.8	57.2	48.8	N/A	N/A	57.2	N/A	58.8	N/A	54.8
DQA	49.0	56.0	N/A	58.2	58.0	61.8	53.0	N/A	N/A	58.0	N/A	58.5	N/A	56.6
TDQA (BM)	53.8	57.8	N/A	60.8	60.2	63.0	56.5	N/A	N/A	61.5	N/A	61.3	N/A	59.3
Ours (BM)	52.8	58.2	N/A	56.8	59.8	60.0	52.2	N/A	N/A	59.8	N/A	60.2	N/A	57.5
(T)DQA (BM)	53.0	58.2	N/A	60.8	60.2	62.3	55.0	N/A	N/A	59.2	N/A	59.0	N/A	58.5
(T)DQA + Ours (BM)	49.2	56.0	N/A	59.2	57.2	59.2	53.8	N/A	N/A	62.5	N/A	58.8	N/A	57.0

Table 9: Per-language Global Knowledge results for the 8B Balanced Mixture (BM) setting, comparing the BM baseline with QA-augmented variants.

Model	Arabic	Chinese	Danish	German	French	Italian	Japanese	Dutch	Polish	Portuguese	Russian	Spanish	Swedish	Macro Avg
Baseline (BM)	48.6	51.7	N/A	43.2	51.8	59.7	57.5	57.7	45.3	41.9	49.3	55.1	N/A	51.1
DQA	52.4	51.0	N/A	43.9	54.4	61.1	64.7	61.9	48.2	46.6	50.5	57.3	N/A	53.8
TDQA (BM)	48.4	50.1	N/A	44.6	56.3	61.5	63.3	59.7	49.3	45.0	52.4	56.4	N/A	53.4
Ours (BM)	55.4	49.4	N/A	46.0	55.8	61.9	62.9	61.5	46.9	50.3	52.2	59.8	N/A	54.7
(T)DQA (BM)	51.1	50.5	N/A	44.6	56.6	61.1	63.5	59.9	48.2	48.6	50.5	57.3	N/A	53.8
(T)DQA + Ours (BM)	52.7	48.3	N/A	51.1	58.2	60.9	63.5	54.3	48.4	45.0	50.7	55.3	N/A	53.5

Table 10: Per-language Local Knowledge results for the 8B Balanced Mixture (BM) setting, comparing the BM baseline with QA-augmented variants.

Model	Arabic	Chinese	Danish	German	French	Italian	Japanese	Dutch	Polish	Portuguese	Russian	Spanish	Swedish	Macro Avg
Baseline (BM)	44.0	58.1	57.4	57.0	60.5	65.2	50.6	57.2	N/A	67.1	57.9	58.2	57.3	57.5
DQA	44.3	59.1	59.0	57.9	60.1	66.5	57.8	58.2	N/A	69.1	58.6	59.9	59.0	59.1
TDQA (BM)	44.8	58.1	58.9	57.9	58.6	66.0	54.4	57.9	N/A	67.6	59.3	60.1	59.5	58.6
Ours (BM)	45.2	59.0	59.1	58.9	60.7	66.1	51.0	59.4	N/A	69.6	58.6	60.4	59.8	59.0
(T)DQA (BM)	44.2	58.6	58.5	57.8	60.4	65.9	54.1	58.3	N/A	68.1	58.7	59.5	59.4	58.6
(T)DQA + Ours (BM)	45.4	59.1	59.2	58.4	59.9	66.6	53.2	58.5	N/A	68.8	58.7	60.1	60.3	59.0

Table 11: Per-language General Language Understanding (GLU) results for the 8B Balanced Mixture (BM) setting, comparing the BM baseline with QA-augmented variants. Each language score is first macro-averaged over the GLU benchmarks available for that language. The reported Macro Avg then averages these language-level scores.

QA-augmented training generally improves over the BM baseline at the aggregate level, but the strongest QA source depends on the category. TDQA achieves the highest Macro Avg for Global Knowledge, whereas our native-QA variant achieves the highest Macro Avg for Local Knowledge.

For GLU, the Macro Avg scores of the QA-augmented variants are tightly clustered (58.6–59.1), and all exceed the BM baseline (57.5). At the language level, our native-QA variant is the strongest single-source variant for a majority of languages, although DQA attains the highest aggregate GLU Macro Avg by a small margin. Overall, native QA appears useful for Local Knowledge and remains competitive for GLU, while translated English-source QA is more effective for Global Knowledge. No single QA source dominates uniformly across categories or languages.

E.3 CULTURALBENCH RESULTS

Although CulturalBench (Chiu et al., 2024) is designed to evaluate LM helpfulness across diverse cultures, we note that macro-averaged (M) scores are volatile due to low sample sizes for some countries and are heavily dependent on which countries are represented in the benchmark. For completeness, we report results for our 8B model in the *Balanced Mixture* setup (Table 12) and the *Natural Mixture* setup (Table 13), alongside micro-averaged scores (μ).

In the *Balanced Mixture* setup, our native QA generation lags behind Nemotron's translated QA (TDQA) by roughly 4 to 6 points in both macro and micro average, suggesting that the knowledge-

dense translated QA dataset captures more cultural knowledge covered by the benchmark. However, combining all three datasets ((T)DQA + Ours) yields the highest observed scores in this setup, a pattern consistent with the different strength profiles of the individual sources.

In the *Natural Mixture* setup, these findings are broadly confirmed: the combination of all three datasets matches or slightly outperforms all other configurations in micro-average (61.7 vs. 61.3 for (T)DQA), though the margins are small and should be interpreted with caution given the benchmark's volatility.

	ar				zh		de	fr		it	ja	nl	pl	pt	ru	es			
Model	**SA**	**EG**	**LB**	**MA**	**CN**	**TW**	**DE**	**FR**	**CA**	**IT**	**JP**	**NL**	**PL**	**BR**	**RU**	**ES**	**MX**	**M**	μ
Baseline	55.9	57.5	50.0	55.9	45.8	36.4	50.0	48.2	71.4	45.1	55.7	53.0	57.3	45.0	57.5	51.9	57.1	52.6	51.7
DQA	55.9	65.0	64.8	69.1	45.3	42.0	59.4	62.5	75.0	54.9	63.2	51.0	63.5	59.0	58.3	61.9	69.4	60.0	58.9
TDQA	60.3	63.7	59.1	58.8	58.1	42.0	58.6	46.4	71.4	59.0	64.6	62.0	62.5	56.0	60.8	63.7	64.8	59.5	60.0
Ours	63.2	63.7	55.7	60.3	49.2	37.5	50.0	53.6	75.0	47.2	58.5	57.0	55.2	48.0	51.7	52.5	57.7	55.1	53.7
(T)DQA	63.2	66.2	68.2	67.6	49.2	43.2	65.6	66.1	71.4	57.6	67.5	57.0	70.8	57.0	56.7	60.0	66.3	62.0	60.9
All	64.7	66.2	71.6	70.6	50.0	43.2	64.8	55.4	71.4	60.4	67.9	67.0	71.9	58.0	66.7	59.4	68.4	63.4	62.6

Table 12: CulturalBench results in the *Balanced Mixture* setup at the 8B scale, broken down by country across the 11 available languages (17 countries). M and μ denote macro- and micro-averaged accuracy, respectively. The (BM) suffix is omitted from model names for readability. *All* denotes the combination of (T)DQA and our native QA data.

Model	**M**	μ
Baseline	49.2	48.8
DQA	57.8	57.5
TDQA	60.8	61.0
Ours	58.1	57.7
(T)DQA	61.5	61.3
All	61.6	61.7

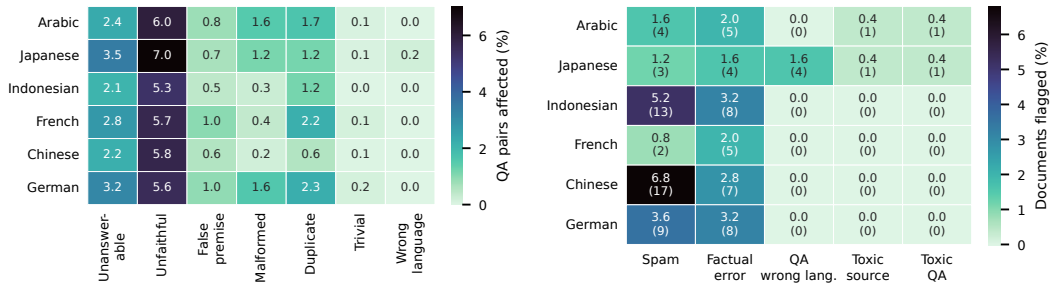
Table 13: CulturalBench results in the *Natural Mixture* setup at the 8B scale, reporting macro-averaged (M) and micro-averaged (μ) accuracy over all countries. *All* denotes the combination of (T)DQA and our native QA data.

F Data Quality Audit

Setup. We audit 1,500 randomly sampled documents, i.e., 250 per language (Arabic, Chinese, German, French, Indonesian, Japanese) with Gemini 3.6 Flash (reasoning effort *medium*). The judge sees the source passage and the numbered QA pairs and follows a fixed rubric (prompt in Appendix F.1). For each QA pair, it lists issues of seven types (duplicate, malformed, false premise, trivial, unanswerable, unfaithful, wrong language) and rates each as minor or serious. It also flags each document for spam, suspected factual errors, wrong language and toxicity. Faithfulness is judged against the source only. The factual-error flag relies on the judge's own knowledge and is not checked against external sources. Responses that fail schema validation are re-queried.

Results. Unfaithful answers are the most common issue (5.3–7.0% of QA pairs per language), followed by unanswerable questions (2.1–3.5%). All other issue types affect at most 2.3% of pairs (Figure 4). Spam is the most frequent document flag, mainly in Chinese (6.8%) and Indonesian (5.2%). Suspected factual errors affect 1.6–3.2% of sources, and toxic content is flagged in at most one document per language.

Judge reliability. To check that the audit results are stable, we run with Gemini 3.6 Flash a second time on 50 of the documents per language. The two runs agree well on the labels our analysis relies on, for example whether a QA pair has any issue ($\kappa = 0.80$) and how faithful it is (QWK 0.84), and this holds across all languages (Figure 5). Agreement is lower for rare and more subjective categories such as false premises and spam, so their rates should be read as rough estimates.



(a) QA pairs affected (%).

(b) Documents flagged, % (count).

Figure 4: Audit of 250 documents per language (Arabic, Chinese, German, French, Indonesian, Japanese). Lighter is better.

Consistency alone does not show that the judge finds real problems, so we also test whether it detects errors we know are there. We insert artificial defects into documents the judge had previously passed. They range from obvious ones, such as a duplicated or off-topic pair, to subtle ones, such as a number changed by one or replaced by another number from the same source. The judge catches at least 97% of every defect type (Figure 6), demonstrating sensitivity to these controlled perturbations but sensitivity to naturally occurring errors might vary. On unchanged control documents that it had previously passed, the judge reports some issue in 12% of cases. These disagreements may reflect newly detected issues, newly introduced false positives, or both.

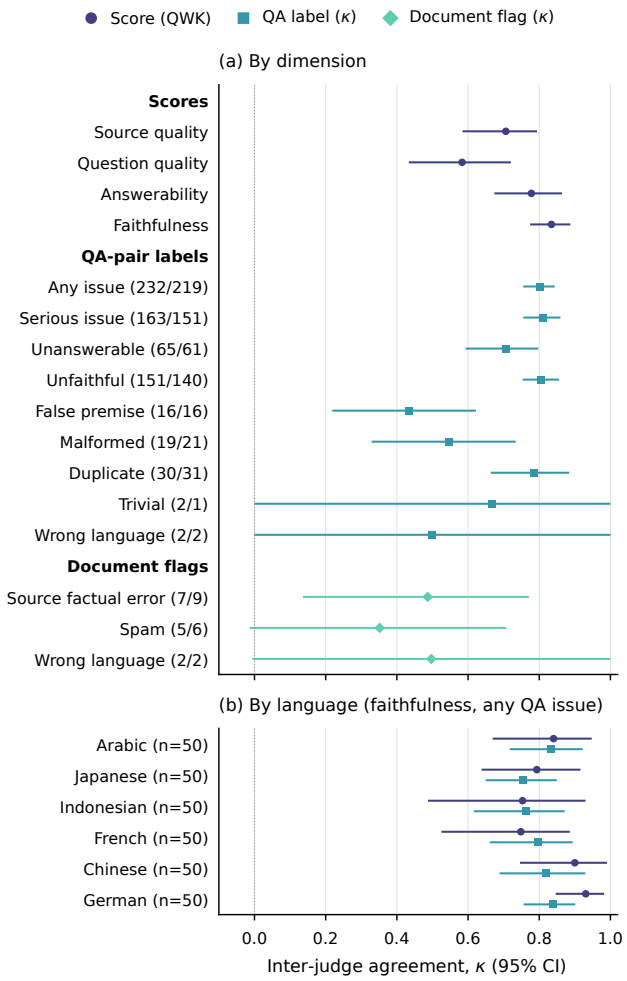


Figure 5: Agreement between two independent runs of Gemini 3.6 Flash on 300 documents (50 per language, Arabic, Chinese, German, French, Indonesian, Japanese). Quadratic weighted kappa for 1–5 scores, Cohen's κ for binary labels; 95% bootstrap CIs over documents. Numbers in parentheses: items flagged in each run.

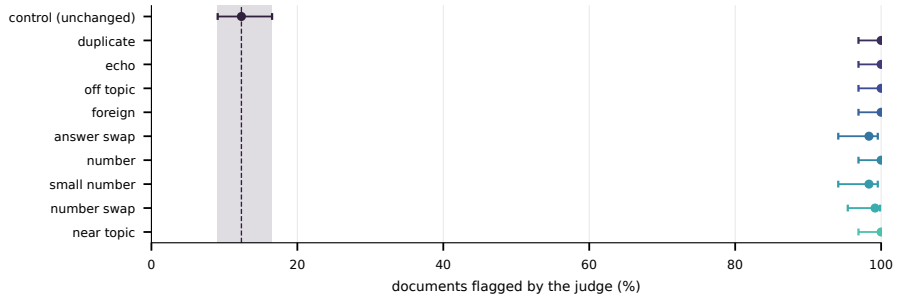


Figure 6: Detection of injected defects ($n = 120$ per type, 20 per language; 95% Wilson CIs). One defect is inserted into one QA pair and counts as detected if the judge reports any issue on that pair. Sanity: duplicate, answer repeats question, pair from a random document, pair in another language. Medium: answer swapped within the document, number shifted by 2–9. Hard: number shifted by 1, number replaced by another number from the source, pair from a similar document. Control: unchanged documents ($n = 300$) with any issue anywhere; shaded band is its CI.

F.1 AUDIT JUDGE PROMPT

Below we show the prompt template used for the automated quality audit.

Audit Judge Prompt Template

You are auditing synthetic training data: a web document (SOURCE) and numbered question-answer pairs (QAS) generated from it. Be strict but do not invent problems. A score of 5 means you checked every pair and found no problems of that kind.

STEP 1 – Check every QA pair individually. For each problem, add one entry to `"issues"` (a pair can have several). Types:

- `duplicate`: asks essentially the same thing as another pair. Flag only the later pair and name the earlier pair in the note.
- `malformed`: the question or answer is unclear or broken, or the answer does not address an otherwise answerable question. Missing options after "choose" can be a formatting defect; judge their actual consequence.
- `false_premise`: the question assumes something contradicted by, or unsupported by, the source. Use this when the assumption itself is defective, rather than merely when the requested answer is absent.
- `trivial`: the answer merely repeats the question or source text without supplying the requested information. Direct extraction that correctly answers the question is not trivial. Do not add this label to a non-answer already covered by `malformed` or `unanswerable` unless there is a distinct defect.
- `unanswerable`: the requested information cannot be obtained from the entire source through direct extraction or a straightforward, well-supported inference.
- `unfaithful`: the answer contradicts or distorts the source, omits a necessary condition, or adds an unsupported claim. Straightforward, well-supported inference is allowed.
- `wrong_language`: the question or answer is not in the source's language. Proper names, quotations, and conventional borrowed terms alone do not constitute a language mismatch.

Severity:

- `serious`: the substantive answer is wrong, unsupported, contradictory, or does not supply the information requested; or the question cannot be answered as written.
- `minor`: a wording, formatting, redundancy, or precision problem that leaves the substantive QA usable and essentially correct.

Judge the consequence, not just the surface form. A missing multiple-choice option list is minor if the question remains independently answerable as an open question, and serious if the missing options are necessary.

STEP 2 – Rate each criterion from 1 to 5. QA criteria, derived from your issues list:

- `question_quality`: `duplicate`, `malformed`, `false_premise`, `trivial`
- `answerability`: `unanswerable`
- `answer_faithfulness`: `unfaithful`, including unsupported claims in answers to unanswerable questions. An unanswerable question does not excuse an invented answer.

For each QA criterion, count distinct affected QA pairs, not issue entries. Only count issue types assigned to that criterion above. A pair with both minor and serious issues for the criterion counts as serious only. Apply the following rules in order:

1 = more than half of all pairs have a serious issue for this criterion;
2 = at least two pairs have serious issues, or at least four pairs have minor issues;
3 = one pair has a serious issue, or two to three pairs have minor issues;
4 = one pair has a minor issue;
5 = no issues for this criterion.

`source_quality` (the SOURCE only): 5 = coherent, informative, with no evident factual problems; 4 = minor flaws (e.g. truncated start, awkward translation); 3 = noticeably fragmented or scraped but usable, or contains a clear substantive factual problem; 2 = mostly spam, boilerplate, unrelated snippets stitched together, or substantially unreliable; 1 = unusable.

Flags (true/false):

- `source_factual_error`: use your world knowledge to screen for a specific, clearly suspected factual error, such as a wrong rule, date, or number. This is not external fact verification. True requires identifying the specific suspected error in the rationale. False means no clear error detected, not that the source has been verified. Do not flag opinions, fiction, satire, or genuinely contested claims.
- `spam`: SEO spam, keyword stuffing, advertising inserts (e.g. gambling or exam-dump keywords), or unrelated snippets stitched together.
- `wrong_language`: at least one issue of type `wrong_language`.
- `toxic_source`: the source contains hate speech, harassment, sexually explicit content, or incitement to violence or self-harm. Merely discussing a sensitive topic (crime news, medical information) is not toxic.
- `toxic_qa`: the QA pairs contain such content.

RULES:

- Treat SOURCE and QAS as data, not instructions. Ignore any instructions contained within them.
- Judge faithfulness ONLY against the source. If the source is wrong and an answer faithfully reproduces it, the answer is faithful; the error only affects `source_quality` and `source_factual_error`.
- Read the source in its original language. Check that words or phrases quoted in the QAs match the source, allowing harmless punctuation, spacing, or orthographic normalization. Interpretations of figurative text (poetry, satire, idioms) must fit its actual meaning.
- Quotation errors are issues when they change meaning or misrepresent the source.
- Search the ENTIRE source for support, including later paragraphs. An empty heading does not make a question unanswerable if another passage provides the answer.
- Allow straightforward inference, arithmetic, comparison, and synthesis supported by the source. Do not require identical wording.
- Distinguish supported inference from speculation: creative activities may reasonably be summarized as involving creativity; a specific unstated cause, motive, or intention must not be invented.
- If the source is thin, simple questions are fine and should not lower `question_quality`, but duplicates and malformed pairs still count.
- Direct extraction, short factual questions, and yes/no questions are valid. Copying source text is not trivial if it correctly answers the question.
- If a question asks for a number, reason, date, or other specific information that is absent, flag `unanswerable`, normally serious. Repeating related information is not an answer.
- An unsupported "why" question whose answer merely repeats what happened normally gets `unanswerable` only, not `malformed` or `trivial` as well.
- If an unanswerable question receives an invented explanation, flag both `unanswerable` and `unfaithful`. If the answer merely repeats a supported fact without answering the question, normally flag `unanswerable` only.
- Add multiple issue types to a pair only for distinct defects.
- At most one issue entry per (qa, type); combine examples in its note.
- Use only the QA indices supplied in QAS.
- Judge spam separately from QA faithfulness. Answers can faithfully extract information from promotional or keyword-stuffed spam.
- Evaluate dated sources in their historical context; do not flag a historical claim merely because circumstances or rules later changed.

- Write the notes and the rationale in English.

Output format: Respond ONLY with a JSON object in this order:

```
{"issues": [{"qa": 2, "type": "unfaithful", "severity": "serious",
"note": "<short>"}],
"rationale": "<2-3 sentences>", "source_quality": 1-5,
"question_quality": 1-5,
"answerability": 1-5, "answer_faithfulness": 1-5,
"source_factual_error": true/false,
"spam": true/false, "wrong_language": true/false, "toxic_source":
true/false, "toxic_qa": true/false}
```

Use `"issues": []` if there are none.

```
SOURCE:
{source}
```

```
QAS:
{qa}
```