Scaling Effects on Multilingual Performance

Xinhe Shi[1], Qingcheng Zeng[1]*, Weihao Xuan[2]*, Kaize Ding[1]†,

[1]Northwestern University, [2]The University of Tokyo
shixinhe2024@outlook.com, qcz@u.northwestern.edu,
weihaoxuan@g.ecc.u-tokyo.ac.jp, kaize.ding@northwestern.edu

Abstract

Small-scale language models are increasingly prioritized in real-world applications due to their low latency, reduced energy consumption, and modest computational requirements. However, it remains underexplored ***whether model performance on different languages degrades at a comparable rate under model downsizing***. This question is critical for assessing the viability of compact models in equitable multilingual applications. Intuitively, the imbalanced training corpora may cause model performance on low-resource languages to deteriorate faster, undermining compact models' usability in global deployments. To investigate this, we establish the **first comprehensive hierarchical task taxonomy** and then introduce a Massive Multilingual Multitask Multiple Choice Question dataset (**M4CQ**), which covers **119** tasks across various domains and **19** languages, with 55673 manually reviewed high-quality questions per language. We conduct experiments on over ten models of various series and sizes, and results show that the resource richness of languages does not have a stable and significant impact on the scaling effect of multilingual performance, indicating that compact models can preserve relative multilingual competence. Our systematic investigation not only **highlights the potential of small-scale models for multilingual applications**, but also **contributes to the advancement of global AI equality**. The dataset is available at https://huggingface.co/datasets/LearnerSXH/M4CQ.

1 Introduction

Large language models (LLMs) have demonstrated unprecedented multilingual capabilities, yet their deployment remains constrained by latency, energy cost, and computational demands. To bridge this gap, the community has actively explored smaller models through distillation, pruning, and parameter-efficient tuning. While prior work shows that smaller model size typically leads to poorer performance, it still remains unclear **whether all languages suffer equally as models shrink in size**. Specifically, when multilingual ability diminishes with reduced model capacity, will low-resource languages deteriorate more rapidly than high-resource ones?

In multilingual settings, models are typically trained on highly imbalanced corpora, where high-resource languages dominate both in data volume and syntactic–semantic coverage. Consequently, when model size shrinks, the already limited representational capacity may be disproportionately allocated to frequent, high-resource languages, potentially accelerating performance collapse in low-resource ones. Therefore, intuitively, model performance on low-resource languages may degrade faster when the model size is reduced.

This asymmetry has further implications:

- if low-resource languages suffer significantly faster degradation than high-resource counterparts as models shrink, it would suggest *a fundamental limitation of small-scale models in achieving equitable multilingual performance*, undermining their viability for truly global deployment and highlighting the need for linguistic-equal compression strategies in future multilingual model design;

- conversely, if performance degradation proceeds at a comparable rate across high- and low-resource languages, it would indicate that *compact models can preserve relative multilingual competence*, thereby supporting their use in inclusive, multilingual applications (highlighting their viability for globally inclusive deployment and motivating linguistically fair compression strategies).

From a mechanistic perspective, these two possible outcomes may stem from fundamentally dif-

*Co-second authors
†Corresponding author

Figure 1: The hierarchical task categorization diagram.

ferent properties of multilingual representations. If low-resource languages indeed deteriorate faster under model downsizing, a plausible explanation is that their representations rely more heavily on excess model capacity and implicit parameter redundancy, as they are primarily learned through indirect transfer and parameter sharing rather than abundant language-specific evidence. When capacity is reduced, such fragile and sparsely grounded representations may be pruned or overwritten earlier during optimization, leading to disproportionate degradation. In contrast, if performance declines at similar rates across languages, it would suggest that multilingual models learn shared, well-aligned representations that are largely independent of scale, where cross-lingual abstractions rather than language-specific memorization dominate. In this case, even compact models may retain robust multilingual competence, as core linguistic structures are preserved through efficient parameter sharing and scale-invariant alignment. Understanding which of these mechanisms governs multilingual scaling behavior is therefore crucial for both theoretical insight and practical multilingual model design.

In order to investigate this research question, it is essential to comprehensively evaluate the model's performance across various aspects on each language. However, existing multilingual benchmarks such as Global-MMLU (Singh et al., 2024) and MMLU-ProX (Xuan et al., 2025) suffer from a key limitation: although these benchmarks are indeed multilingual, they cover only a single task type (knowledge-based tasks), which is insufficient for evaluating a model's abilities within any given language. For instance, a model that performs well on Arabic mathematics questions does not necessarily excel at Arabic tasks of other task types, such as coreference resolution. One might attempt to mitigate this limitation by jointly utilizing multiple multilingual benchmarks which targets different task types. Nevertheless, such an approach remains fundamentally constrained. First, it restricts the set of analyzable languages to the intersection of languages supported by all selected benchmarks, substantially limiting the scope of multilingual analysis. Second, task diversity obtained through enumerating specific multilingual benchmarks is inherently ad hoc and incomplete, as the chosen tasks reflect dataset availability rather than a principled or comprehensive coverage of linguistic and reasoning skills. As a result, current evaluation practices fall short of supporting systematic analysis of multilingual performance variation across languages and model scales. Therefore, what we truly need is a multilingual, multi-task benchmark that covers as diverse task types as possible while maintaining content consistency across languages, enabling fair evaluation of performance in every language.

To address this, we establish **the first hierarchical task categorization diagram** (Figure 1) through systematic survey of more than 20 existing datasets, providing a unified framework for task categorization. Building upon this taxonomy, we introduce a Massive Multilingual Multitask Multiple Choice Question (**M4CQ**) dataset which features domain-diverse distribution of **119** tasks and content consistency across **19** languages. The M4CQ dataset enables systematic and consistent evaluation of models across diverse languages, making it possible to investigate patterns of multilingual capability degradation as model size decreases, and to implement our proposed approach for ranking the resource richness of different languages.

Through experiments on over ten models of various series and sizes, and results show that **the resource richness of languages does not have a stable and significant impact on the scaling effect of multilingual performance**, indicating that compact models can preserve relative multilingual competence.

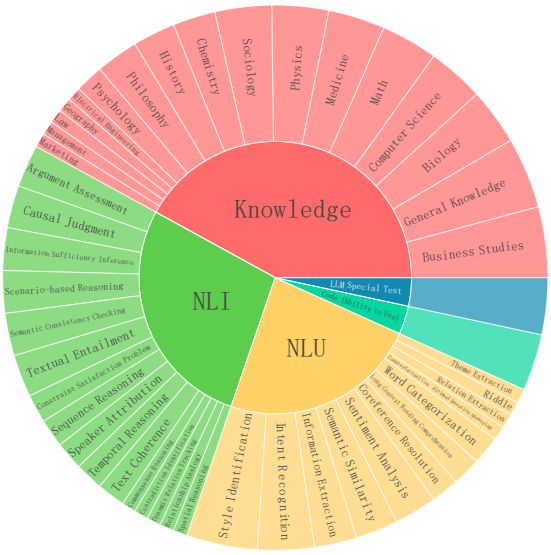

Figure 2: Domain distribution of 119 tasks in M4CQ, where each domain's angular span represents the proportion of its constituent tasks.

Code	Language	Code	Language
AR	Arabic	BG	Bulgarian
CS	Czech	EL	Greek
EN	English	FI	Finnish
HU	Hungarian	ID	Indonesian
IT	Italian	JA	Japanese
LT	Lithuanian	LV	Latvian
NL	Dutch	RO	Romanian
RU	Russian	SK	Slovak
TR	Turkish	UK	Ukrainian
ZH	Chinese (Simplified)		

Table 1: Language Codes and Full Names in M4CQ Dataset

2 The M4CQ Dataset

Given the absence of an existing open-source dataset that meets our requirements, we constructed the Massive Multilingual Multitask Multiple Choice Question (**M4CQ**) dataset to assess LLMs' capability comprehensively on 19 languages.

2.1 Dataset Overview

The M4CQ dataset consists of **19 languages** (covering high-/low-resource languages), each containing **119 task categories** in multiple-choice format. Figure 2 shows M4CQ's task distribution, demonstrating a balanced spread across various domains. Each instance has semantically equivalent versions in all languages. See Appendix A for detailed information about these 119 tasks and examples in M4CQ. Table 1 provides a complete list of languages covered in M4CQ.

M4CQ contains five macro-categories: Knowledge, Natural Language Understanding (NLU), Natural Language Inference (NLI), Code (Ability to Use), and LLM Special Test. Knowledge covers questions that primarily evaluate factual or domain-specific knowledge across different disciplines. NLU includes tasks that can be answered through fundamental language understanding, such as grasping surface meaning, semantics, or straightforward textual information. NLI refers to tasks that go beyond basic comprehension and require more advanced reasoning, inference, or logical deduction based on the given text. Code comprises tasks that involve programming-related content, including code comprehension, generation, or reasoning about code behavior. Finally, LLM Special Test denotes task types that are explicitly designed for model-centric evaluation, and will not be used to assess the abilities of real persons.

A detailed comparison between M4CQ and other existing multilingual multitask datasets is shown in Appendix B.

2.2 Construction Methodology

2.2.1 Summary of Hierarchical Task Categorization

As mentioned above, we need a multitask dataset covering comprehensive domains to better assess models' abilities on each language. However, existing datasets focus either on a single task type or on a group of isolated tasks, lacking a comprehensive coverage of task domains. Moreover, there is no systematic categorization or structured summary of task domains in prior work.

To address this, we established the first hierarchical task categorization diagram (Figure 1) through systematic survey of existing datasets (Johannes Welbl, 2017; Williams et al., 2018; Clark et al., 2018, 2019; Bhagavatula et al., 2020; Wang et al., 2018; Mihaylov et al., 2018; Talmor et al., 2019; Zellers et al., 2019; Wang et al., 2019; Bisk et al., 2020; Hendrycks et al., 2020; Liu et al., 2020; Sakaguchi et al., 2021; Chen et al., 2021; Cobbe et al., 2021; Srivastava et al., 2022; Rein et al., 2023; Zhong et al., 2023; Parvesh and at XenArcAI, 2024; Liu et al., 2024), providing a unified framework for task categorization.

Building upon this taxonomy, the dataset we de-

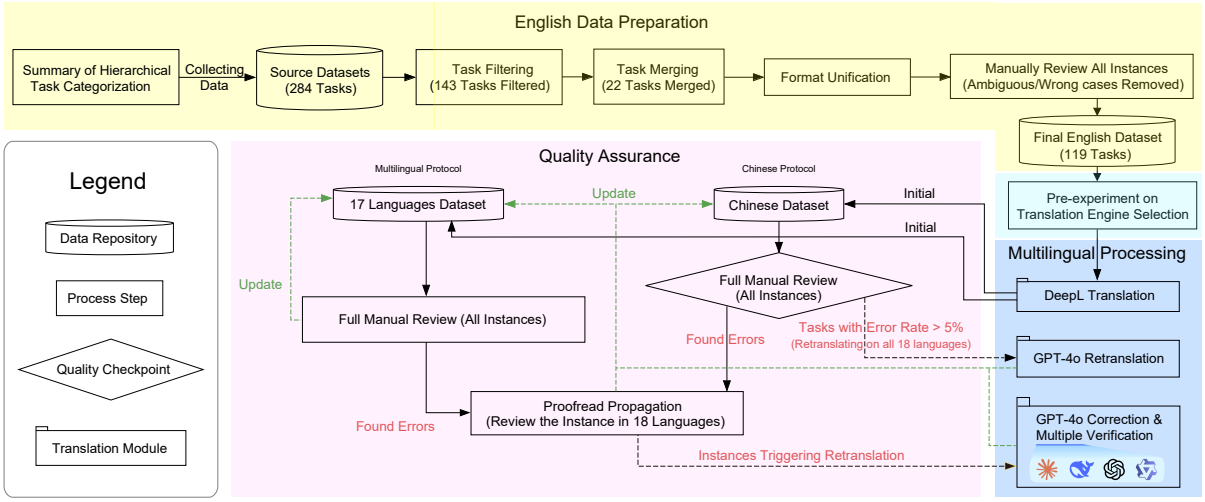

Figure 3: The construction workflow of M4CQ. The workflow begins with the creation and multi-step processing of an English dataset, which ensures the high-quality of source data. Then it's translated into 18 languages using DeepL, a translation engine standing out in our pre-experiment (detailed in Appendix D). To guarantee the quality of non-English datasets, non-English instances are manually verified by proficient annotators, with flawed instances checked in all 18 languages (called as "proofread propagation").

veloped spans the full spectrum of NLP tasks, not only meeting the requirements of our experiments but also offering a holistic benchmark for evaluating model capability across domains.

2.2.2 Task Curation

1. Task Filtering: a) Removed language-biased tasks like code reasoning (biased towards English, but pseudo-code is acceptable), English proverbs, U.S. history questions. b) Excluded low-quality tasks where answer selection of questions is unconvincing. 2. Task Merging: Consolidated tasks of the same category across sources (e.g., combining BIG-bench's *presuppositions as nli* with GLUE's *MNLI*). 3. Format Unification: Converted to a format supporting flexible option counts and future extension to multi-answer questions (detailed in Appendix C), which overcomes the rigidity of fixed-option formats (e.g., "A/B/C/D" constraints). 4. Instance Filtering: Manually reviewed all instances and removed instances where answer selection is wrong or unconvincing.

2.2.3 Cross-Lingual Alignment and Quality Control

Note: Due to the significantly higher availability of high-caliber Chinese(zh)–English(en) bilingual experts for us compared to those in other language pairs, we differentiated the quality control pipeline for Chinese and other 17 target languages. And the latter were manually proofread through round-trip translation en→L_x→zh→en, which ensured the validation quality since wrong middle results could not be translated into correct ones any more especially when there're two middle languages.

Translation Pipeline and Validation for Chinese 1. Select Translation Engines: We conducted a preliminary experiment to select suitable translation engines, where DeepL[1] showed great accuracy and stability, contributing to its role as primary engine. Appendix D provides details about the pre-experiment. 2. Initial Translation: All tasks were translated from English to 18 target languages using DeepL at the instance level (question + choices). 3. Full Verification for Chinese: Every instance of the 119 tasks in Chinese was manually reviewed and corrected by native Chinese speakers proficient in English. For the identified mistranslation instances, we labeled these as "error-prone instances" (which might contain elements hard to translate and lead to same flaws in translation to other languages), and therefore performed Proofread Propagation—reviewing its translations in all 17 other languages. 4. Identification of DeepL Limitations: During the verification process above, we observed that DeepL underperformed on 9 tasks (e.g., college computer science, abstract algebra) with error rate >5%. These tasks were re-translated into all 18 languages using GPT-4o (OpenAI et al., 2024) and underwent the same manual review process as above.

[1] https://www.deepl.com

During the validation process for Chinese, we noticed that tasks can be divided into "error-prone" category (only the nine tasks mentioned above) and "non-error-prone" category. For the former, DeepL showed high error rate, while for the latter, DeepL was almost error-free. So we designed sampling plus threshold check as validation for non-Chinese languages.

Validation for Non-Chinese Languages 1. For each non-Chinese target language L_x: a) Randomly sampled 30% instances per task. b) Performed round-trip translation: $L_x \rightarrow$ zh $\rightarrow$ en. c) Compared the result with the original English version. d) Flagged instances where semantic inconsistency occurred. e) Corrected flagged instances under multi-verification (using Claude 3.5 Sonnet, DeepSeek-R1 (DeepSeek-AI et al., 2025), GPT-4o, and Qwen2.5-72B-Instruct (Qwen et al., 2025) in collaboration) and reviewed corresponding instances in other 16 languages (Proofread Propagation). 2. Error Thresholding: If samples in any task in L_x had >3 flagged instances, we conducted full manual review of all instances in that task of language L_x. Considering our tasks' size, this threshold is small enough.

2.3 Potential Applications

The M4CQ dataset's unique design (see Appendix B for comparison between M4CQ and other existing multilingual multitask datasets) enables broad applications beyond our research. Here we highlight some directions. **Under-Resourced Language Understanding**: M4CQ provides high-quality, translation-verified data for 10+ medium- and low-resource languages (e.g., Lithuanian, Latvian, Slovak), enabling research on language-specific pre-training and data augmentation (e.g., back-translation, synthetic data generation) for under-resourced languages. **Multilingual Model Benchmarking**: With balanced task distribution and strict semantic equivalence across languages, M4CQ can serve as a benchmark for evaluating Multilingual Language Models, allowing direct comparison of model performance across different languages. **Multitask Learning Optimization**: The 119 task categories spanning diverse domains (Figure 2) enable research on multitask learning optimization, including task weighting strategies and gradient conflict resolution, while supporting analysis of inter-task correlations and their impact on model performance.

These applications highlight M4CQ's value as a versatile resource for multilingual NLP research. Its strict cross-lingual alignment and task neutrality uniquely support disentangling linguistic capabilities from task-specific biases.

3 Experimental Setup

To systematically investigate the scaling effects on multilingual performance, we conduct extensive evaluations across diverse model families and language scales using our proposed M4CQ benchmark.

3.1 Model Selection

To ensure the generalizability and robustness of our findings across diverse architectural configurations and pre-training paradigms, we evaluate over ten language models from several model families, including Qwen3 (Yang et al., 2025), Phi-4 (Microsoft et al., 2025), Mistral[2], and Aya (Dang et al., 2024). We adhere to a principled selection process based on these key criteria:

1. **Industry Prominence:** We prioritize models that are widely adopted within the research community and industry, ensuring that our observations reflect the scaling behaviors of state-of-the-art (SOTA) architectures.

2. **Timeliness:** We select the most recent iterations of each model family (e.g., Qwen3 rather than Qwen2) to ensure our analysis accounts for the latest advancements in pre-training techniques and data curation strategies.

3. **Sufficient Scaling Depth:** To reliably measure the performance degradation rate, a model family must exhibit a significant parameter delta. Specifically, we require the difference in the amount of parameters between the largest and smallest models within a family to exceed 10B. This threshold ensures that the observed scaling effects are substantial enough to provide representative and generalizable insights rather than being confounded by minor capacity fluctuations.

For each family, we evaluate all available parameter sizes (from the smallest compressed versions to the largest foundation models) to capture the complete scaling trajectory.

[2] https://huggingface.co/mistralai/models

There are still some model families not selected in our analysis. Due to space constraints, please refer to Appendix E for detailed justifications for the exclusion of specific families.

3.2 Evaluation Framework and Tasks

We utilize the *Harness* framework (Gao et al., 2024) to ensure a standardized and reproducible evaluation pipeline. All models are evaluated on the **M4CQ** dataset with zero-shot settings.

3.3 Evaluation Metrics

For each model M and language L, we use the **Accuracy (Acc)** as the primary metric. We compute:

- **Total Acc:** The mean accuracy across all 119 tasks within language L.
- **Category Acc:** The mean accuracy across tasks within each of the five macro-categories (i.e., Knowledge, NLU, NLI, Code and LLM Special Test).

3.4 Measuring Performance Degradation

To isolate the effect of model scaling from the absolute performance baseline of different languages, we define the **Degradation Rate** (DR). For any given model M_i, let M_{max} be the model with the largest parameter size within the same model family as M_i. The degradation rate for a given accuracy metric Acc (could be Total Acc or any Category Acc) in language L is defined as:

$$DR(M_i, L) = \frac{Acc(M_i, L)}{Acc(M_{max}, L)} \times 100\% \quad (1)$$

By plotting DR against model parameter size (ordered from largest to smallest on the x-axis), we can obtain a line chart and visualize the rate of performance collapse, with each line corresponds to a different language. A steeper slope indicates a higher sensitivity to model shrinking. This normalized metric allows for a direct comparison between different languages, regardless of their initial performance gaps on the M_{max} model.

4 Results and Analysis

4.1 Overall Multilingual Scaling Behaviors

Figure 4 presents the overall multilingual performance degradation trajectories for four representative model families (Qwen3, Phi-4, Mistral and Aya), covering 19 languages across multiple model scales. For each model family, performance is normalized by the largest model in the series, and the resulting Degradation Rate (DR) reflects the relative preservation of overall multilingual competence as model capacity decreases. Due to space limitations, please refer to Appendix F for more experiment results. Several key observations emerge from our analysis:

Universal Performance Degradation under Model Downsizing. As expected, for any given model family, overall performance monotonically declines as model size is reduced, and this pattern holds uniformly across all evaluated languages.

Narrow Variance in Degradation Rates Across Languages. While performance declines for all languages, the **rate of decline is close**. The range of DR between the fastest- and slowest-declining languages within any single model family averages only about **12%**. Given that DR is a ratio relative to the largest model's performance, this corresponds to an even smaller absolute difference in accuracy, often just a few percentage points. This narrow spread indicates that no language exhibits a significantly different sensitivity to model capacity reduction.

Unstable Language-Specific Degradation Patterns. A closer inspection of individual languages further reveals that ***DR* rankings are not stable across model families**. Languages that appear relatively robust in one model series do not consistently maintain this advantage in others. For example, Japanese (JA) achieves the highest degradation rate (i.e., the slowest relative decline) in the Qwen3 series, yet ranks only ninth in the Mistral series and fifteenth (among the bottom five) in the Phi-4 series. Conversely, Indonesian (ID) exhibits the strongest robustness in the Aya models but becomes the fastest-degrading language in the Qwen3 series. Similar inconsistencies are observed for other languages. Hungarian (HU) attains the highest degradation rate in the Phi-4 models and ranks third in Mistral, while becoming the fastest-degrading language in the Aya series. Likewise, Italian (IT), which exhibits the lowest degradation rate in the Phi-4 series, ranks eighth in the Aya models. These cross-model reversals indicate that no language exhibits a stable enough degradation profile.

Impact of Resource Richness: The Case of English and Chinese. Although the preceding anal-

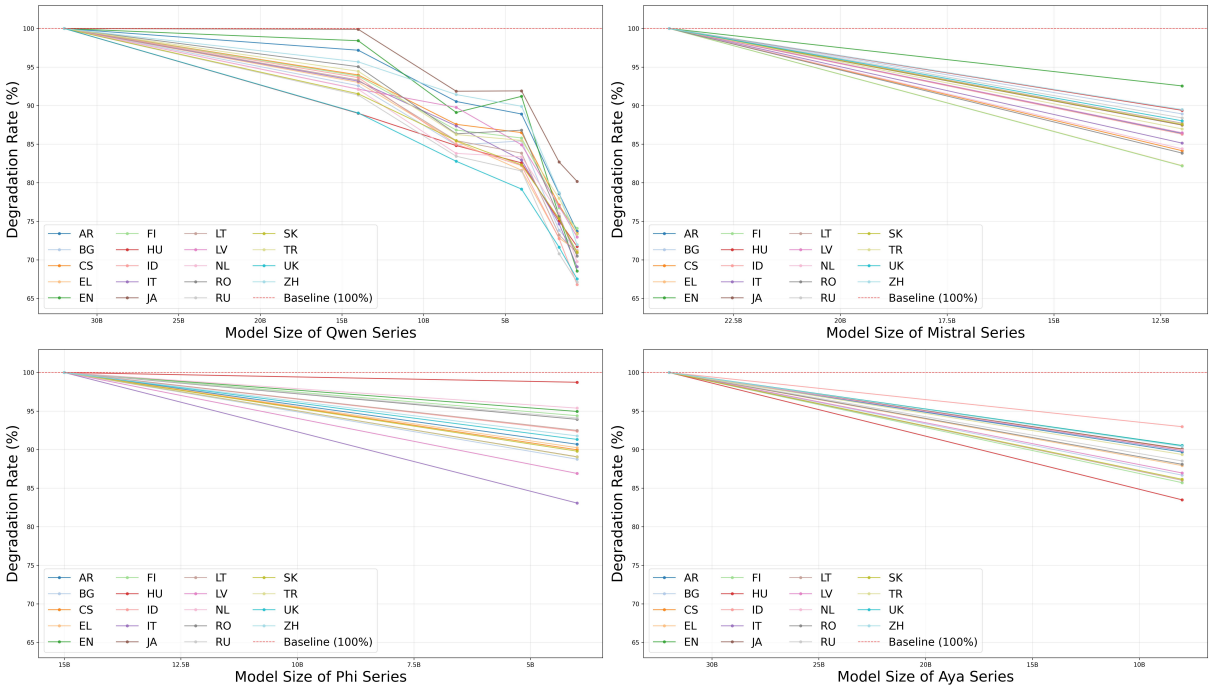

Figure 4: Scaling effects on overall multilingual performance. As model size decreases, performance declines across four model families in all involved languages. However, there is no significant and stable difference in the rate of performance degradation between high-resource and low-resource languages as model size decreases, highlighting their viability for globally inclusive deployment and motivating linguistically fair compression strategies.

yses indicate that no language exhibits a consistently superior or inferior degradation pattern, we further examine whether language resource richness is entirely unrelated to degradation behavior, or whether it exerts a limited or scale-dependent influence. To this end, we focus on two canonical high-resource languages, English (EN) and Chinese (ZH), and analyze their degradation rates across model families and scales.

While English shows a significant drop in Qwen3 when the size is below 4B (ranking 16^{th}), it ranks within the top 6 (among 19 languages) when the model size is larger than 4B. Additionally, Chinese consistently maintains a position in the top half of the DR rankings across all four families (7^{th} in Qwen, 2^{nd} in Mistral, 9^{th} in Phi, and 4^{th} in Aya).

However, while English and Chinese exhibit a slight edge in performance resilience, this advantage is disproportionately marginal when contrasted with their undisputed status as the top two most resource-abundant languages. Their DR rankings are often comparable to, or even surpassed by, much lower-resource languages. Overall, these observations suggest that **language resource richness may correlate with degradation behavior to a limited extent, but does not yield a significant enough impact on the scaling effect of multilingual performance**.

Conclusion on Multilingual Scalability. In summary, our results demonstrate that **there exists no language whose degradation rate is consistently and significantly better or worse than others** across diverse model families. Even high-resource languages like English and Chinese do not exhibit stable, outsized robustness. This result provides strong preliminary evidence against the hypothesis that performance degrades asymmetrically with scaling based on language resource richness.

4.2 Task-Specific Multilingual Scaling Behaviors

We further extend our scaling analysis to the five macro-categories of M4CQ. The DR for each category is computed based on the Category Acc. Figure 5 illustrates the DR trajectories specifically for Knowledge-based tasks. Due to space constraints, more detailed results are provided in Appendix F.

The degradation behaviors in task-wise figures are basically consistent with the overall trends, so similar analyses are omitted. Interestingly, we occasionally observe DR values slightly exceeding 100%, indicating non-monotonic scaling. For

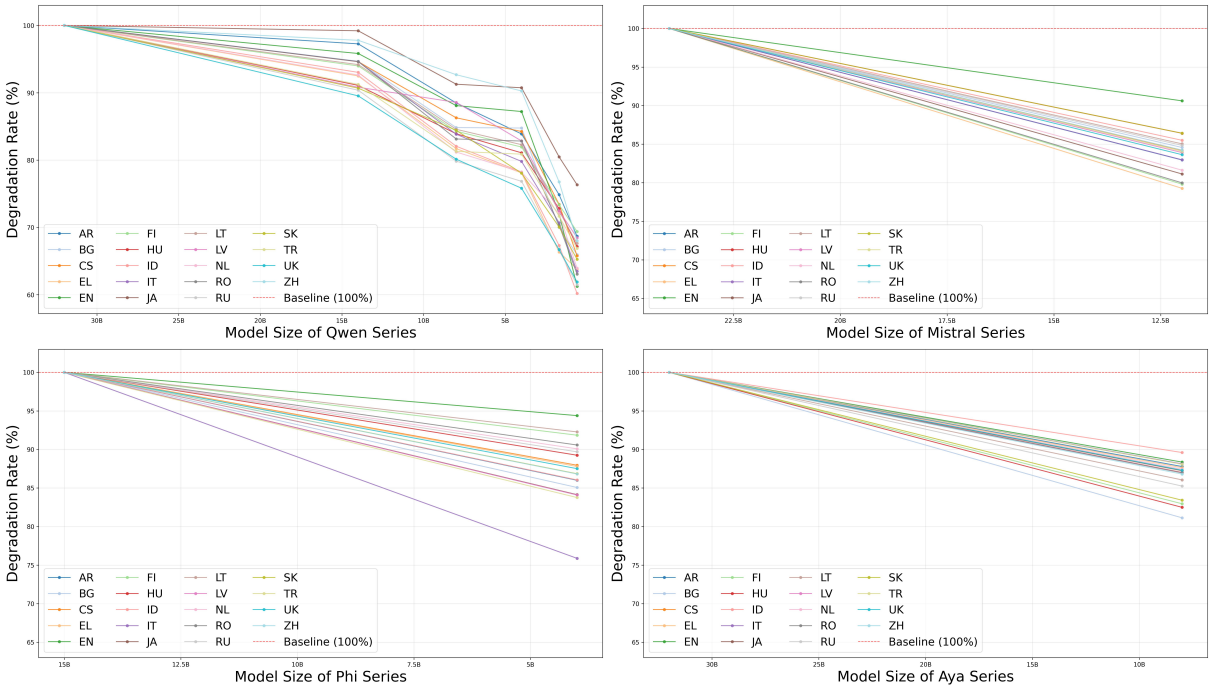

Figure 5: Scaling effects on multilingual performance in knowledge tasks.

low-resource languages such as Dutch and Indonesian, this is likely a numerical artifact where the largest model's poor baseline performance results in a small denominator. In contrast, for high-resource languages like English and Japanese, this phenomenon may suggest parameter saturation: additional parameters yield marginal fluctuations around a performance plateau rather than monotonic gains. Overall, these task-specific observations reinforce the conclusion that scaling effects are largely independent of language resource richness across diverse cognitive dimensions.

5 Discussion

In this work, we set out to investigate whether multilingual performance degradation under model downsizing is asymmetric across languages with different levels of resource richness. Motivated by the highly imbalanced nature of multilingual training corpora, we hypothesized that low-resource languages might deteriorate faster as model capacity is reduced, due to their heavier reliance on indirect transfer and shared parameters. However, our empirical results across over ten models and nineteen languages consistently support the opposite outcome: **performance degradation proceeds at comparable rates across high- and low-resource languages**, and even canonical high-resource languages like English and Chinese do not exhibit a significant advantage in degradation rate.

5.1 Why Do Low-Resource Languages Not Degrade Faster?

One plausible explanation for this finding lies in the nature of multilingual representations learned by large language models. LLMs appear to develop a **shared, universal semantic space** where cross-lingual abstractions dominate over language-specific memorization. Therefore, as the model size shrinks, this shared representational core contracts as a whole, resulting in parallel degradation trajectories across languages. The fact that even high-resource languages like English and Chinese do not exhibit significant advantages implies that their vast data volume primarily improves the absolute performance levels (the task accuracy), but does not fundamentally alter its structural robustness against parameter reduction. This interpretation consists with the observation that DR values of different languages remain close even as their absolute performance levels differ substantially.

5.2 Practical Implications

Our empirical findings provide a strong foundation for the deployment of small-scale models in diverse linguistic contexts. The "Synchronized Degradation" phenomenon is, in fact, an optimistic result for **Linguistic Equity**:

- **Viability of Compact Models for Globally Inclusive Deployment:** Since low-resource languages do not suffer accelerated collapse, compact models (e.g., <10B) remain viable tools for multilingual deployment, preserving the relative multilingual competence of their larger counterparts.
- **Linguistically Fair Compression:** Our work suggests that current compression or scaling strategies have been inherently "fair" in terms of relative performance retention. Developers can pursue model distillation and pruning with greater confidence that these processes will not disproportionately disenfranchise speakers of under-represented languages.

6 Conclusion

In this paper, we present a systematic investigation into the scaling effects on multilingual performance. Our contributions are three-fold:

- We demonstrated that **performance degradation as model size decreases proceeds at a comparable rate across languages of varying resource levels**, refuting the intuitive hypothesis of accelerated collapse for low-resource languages. It indicates that compact models can preserve relative multilingual competence, thereby supporting their viability for globally inclusive deployment and motivating linguistically fair compression strategies.
- We established **the first hierarchical task categorization diagram**, providing a unified framework for task categorization.
- We introduce the **M4CQ** dataset, which features domain-diverse distribution of **119** tasks and content consistency across **19** languages, with 55673 manually reviewed high-quality questions per language.

Limitations

Language Coverage Limitations: While M4CQ includes 19 languages, it still excludes many low-resource and morphologically complex languages critical for comprehensive analysis of language similarity.

Task Coverage Limitations: The 119-task taxonomy, though diverse, may not fully capture all dimensions of linguistic similarity.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 34 others. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

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

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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, and

5 others. 2024. The language model evaluation harness.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Siyao Liu, He Zhu, Jerry Liu, Shulin Xin, Aoyan Li, Rui Long, Li Chen, Jack Yang, Jinxiang Xia, Z. Y. Peng, Shukai Liu, Zhaoxiang Zhang, Ge Zhang, Wenhao Huang, Kai Shen, and Liang Xiang. 2024. Fullstack bench: Evaluating llms as full stack coders. *Preprint*, arXiv:2412.00535.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2024. Mmmlu dataset. https://huggingface.co/datasets/openai/MMMLU. Accessed: 2025-09-14.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Parvesh and Aniket at XenArcAI. 2024. Mathx: Large-scale mathematical reasoning dataset.

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

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Aarohi Srivastava, Abhinav Rastogi, Abhishek B Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 425 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. 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)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yidan Zhang, Yu Wan, Boyi Deng, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, Fei Huang, and Jingren Zhou. 2024. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *Preprint*, arXiv:2411.09116.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Task Information of M4CQ

Table 2 shows domains and descriptions of 119 tasks in M4CQ.

Table 2: Task information of M4CQ

Domain	Task Name	Description
Knowledge - Math	elementary math	Math problems of elementary school difficulty.
	high school math	Math problems of high school difficulty.
	abstract algebra	Questions about abstract algebra.
	college math	Math problems of college difficulty.
Knowledge - Physics	high school physics	Physics problems of high school difficulty.
	college general physics	Examination questions in general physics.
	astrophysics	Questions about astrophysics.
	conceptual physics	Questions about physics concepts.
Knowledge - Biology	high school biology	Biology problems of high school difficulty.
	college biology	Biology problems of college difficulty.
	genetics	Questions about genetics.
	virology	Questions about virology.
Knowledge - Chemistry	periodic elements	Questions about the Periodic Table.
	high school chemistry	Chemistry problems of high school difficulty.
	college chemistry	Chemistry problems of college difficulty.
Knowledge - Philosophy	college philosophy	Examination questions in college philosophy.
	logical fallacies	Questions about logical fallacies.
	moral disputes	Questions about moral disputes.
Knowledge - Sociology	college sociology	Examination questions in college sociology.
	world religions	Questions about world religions.
	security studies	Questions about security studies, related to environmental security, terrorism, weapons of mass destruction, etc.
	public relations	Questions about public relations.
Knowledge - Psychology	high school psychology	Psychology problems of high school difficulty.
	professional psychology	Psychology problems of college difficulty.

Table 2 continued from previous page

Domain	Task Name	Description
Knowledge - Geography	high school geography	Geography problems of high school difficulty.
Knowledge - History	high school world history	World history problems of high school difficulty.
	anachronisms	Given a description, answer if there are any items/phrases that appear out of place in the period context.
	prehistory	Questions about prehistory studies.
Knowledge - Law	international law	Questions about international law.
Knowledge - Medicine	nutrition	Questions about nutrition.
	organology	Questions about the functions of human organs.
	college medicine	Medical problems of college difficulty.
	human sexuality	Questions about human sexuality, related to gender differences, sexual orientation, pregnancy, etc.
Knowledge - Computer Science	high school computer science	Computer science questions of high school difficulty.
	college computer science	Computer science questions of college difficulty.
	computer security	Questions about computer security.
	machine learning	Questions about machine learning.
Knowledge - Electrical Engineering	electrical engineering	Questions about electrical engineering.
Knowledge - Business Studies	business ethics	Business-related gap-fill questions (fill in the blanks by selecting words from the context).
	econometrics	Questions about econometrics.
	high school macroeconomics	Macroeconomics questions of high school difficulty.
	high school microeconomics	Microeconomics questions of high school difficulty.
	professional accounting	Questions about accounting.
Knowledge - Management	management	Questions about management.
Knowledge - Marketing	marketing	Questions about marketing.

Table 2 continued from previous page

Domain	Task Name	Description
Knowledge - General Knowledge	global facts	Questions involving some global statistical data.
	kindergarten knowledge	Kindergarten level general knowledge questions.
	factual judgment	Questions about factual judgment.
	realistic interaction problem	Life-oriented problems that simulate real-world interaction
	scientific common sense	Questions about scientific common sense.
NLU (Natural Language Understanding) - Counterfactual Conditional Question Answering	counterfactualQA	Answer a question based on the given text (inconsistent with facts).
NLU - Information Extraction	sentence info extract	Given a sentence, answer a judgment question based on the sentence.
	table info extract	Given a form, answer a question based on the content of the form.
	context info extract	Answer a question based on a passage (containing 2-8 sentences).
NLU - Long Context Reading Comprehension	gre reading comprehension	Reading comprehension questions in GRE test.
	question selection	Given a context and a short answer (usually a number), choose the corresponding question.
NLU - Coreference Resolution	disambiguation qa	Determine what the pronoun in the sentence refers to in the form of a tautological paraphrase.
	winogrande	The first half of the sentence involves two people, and the second half digs in to ask which one should be filled in.
NLU - Sentiment Analysis	movie review attitude	Given an excerpted sentence from a movie review, evaluate whether its sentiment is positive or negative.
	sentence emo	Determine whether the sentence reveals a positive, negative, neutral or contradictory emotion.
	suicide risk	Given a text, determine the author's suicide risk.

Table 2 continued from previous page

Domain	Task Name	Description
NLU - Semantic Similarity	concept feature	Given a noun phrase, determine which sentence best characterizes the phrase.
	movie recommendation	Choose a movie that is similar to the four movies specified.
	phrase relatedness	Find the most relevant word or phrase.
NLU - Word Categorization	odd one out	Given a set of words, identify the ones that don't fit together.
	commonality abstraction	Find common ground for two nouns.
NLU - Relation Extraction	character relationship	Given a text (basically three or four sentences), determine the character relationship within it.
NLU - Style Identification	authorship identification	Given a text, determine which of the following texts is of the same author as the given text.
	figure identification	Determine the rhetorical device used in the given sentence.
	irony identification	Determine whether the given sentence is ironic.
	snarks	Given two similar sentences, determine which one is ironic.
	humor identification	Determine whether the given text is (cold) humorous.
NLU - Riddle	riddle sense	Brain teaser questions.
NLU - Intent Recognition	goal	Determine the goal of performing an operation.
	step	Answer the steps needed to achieve the given purpose.
	implicatures	Given a conversation in which speaker1 asks a question and speaker2 doesn't answer directly, determine whether speaker2 means yes or no.
	intent	Determining the intent of a sentence.
NLU - Theme Extraction	story moral	Given a story, extract the point the story is trying to make.

Table 2 continued from previous page

Domain	Task Name	Description
NLI (Natural Language Inference) - Textual Entailment	entailment 1p1h2c	Given 1 premise and 1 hypothesis, determine whether it is an entailment (whether the hypothesis can be derived from the premise). 2 choices: entailment/no-entailment.
	entailment fact 1p1h2c	Given 1 fact and 1 hypothesis, determine whether it is an entailment (whether the hypothesis can be derived from the fact). 2 choices: entailment/no-entailment.
	entailment 2p1h2c	Given 2 premises and 1 hypothesis, determine whether it is an entailment (whether the hypothesis can be derived from the premises). 2 choices: entailment/no-entailment.
NLI - Relationship Analogy	similarity analogy	Given examples of six types of similarity, determine which similarity a new example belongs to.
NLI - Text Coherence	logical coherence	Given a text of 10 sentences, it is known that the first few sentences were written by a human and then become computer-written. Find out where the shift begins (it's not a matter of language style, it's that the logic back and forth doesn't make sense anymore).
	content coherence	Given the first half of a paragraph, choose the one that best fits as the next sentence.
NLI - Speaker Attribution	movie dialog same or different	Given an unattributed conversation which comes from a movie, pick out two sentences and ask if they are from the same person.
	play dialog same or different	Given an unattributed conversation which comes from a play, pick out two sentences and ask if they are from the same person.
NLI - Dynamic Relation Tracking	tracking shuffled objects	Given the initial pairing relationship and the subsequent rounds of exchanges, ask what is now the paired item/person for the particular object.

Table 2 continued from previous page

Domain	Task Name	Description
NLI - Spatial Reasoning	navigate	Given a sequence of commands to walk back and forth, determine whether it can get back to the original point.
NLI - Temporal Reasoning	temporal computation	Calculate the date based on the given information.
	temporal sequences	Given someone's schedule for part of the day (the time to do Event A is not mentioned), ask when Event A may be done.
NLI - Sequence Reasoning	order of procedures	Given an objective and two operations, determine in what order the two operations should be performed to achieve that purpose.
	logical sequence	Sorting several objects with intrinsic order relationships.
NLI - Scenario-based Reasoning	scenario qa	Given a scene, ask about feelings/reasons/follow-ups, etc.
	scenario hypothesis qa	Given a scenario, assume one more thing that conflicts with what just happened in the scenario, and ask what might have happened.
	fantasy reasoning	Given a description of a scenario that couldn't happen in reality (e.g. hell, demons, etc.), and ask a judgment question based on it
NLI - Commonsense Reasoning	timedial	Given a piece of dialog in which a word indicating duration/time is obscured, choose a reasonable value for the obscured duration/time.
NLI - Causal Judgment	causal judgment	Given a description of a scenario, judge the causal relationship.
	reasonable causation	Given two sentences in which the causal logic is exactly opposite, determine which sentence contains the correct causal relationship.
	cause extraction	Given a short context, answer the reason (LLM needs to infer from the scenario).

Table 2 continued from previous page

Domain	Task Name	Description
NLI - Information Sufficiency Inference	evaluating information essentiality	Assessing Information Importance: Given a question, determine how useful the following two statements are in answering that question.
	not sufficient	Answer a judgment question based on the given short context. When the information in the context is not sufficient enough to judge, select "Either". (The answer to all instances of this task is "Either".)
	whether sufficient	Given a sentence and a question, determine whether the sentence answer the question.
NLI - Argument Assessment	argument logic	Questions about argument logic.
	logical fallacy detection	Determine whether the given causal logical reasoning is correct or not.
	mathematical induction	Mathematical inference questions.
NLI - Contradiction Identification	lie judgment	Given a short context, identify whether what the character has said is true.
NLI - Semantic Consistency Checking	metaphor understanding	Given a sentence that uses metaphorical rhetoric and an explanation of the metaphorical sentence, answer if this explanation conforms to the meaning of the original metaphorical sentence.
	sentence equivalence	Given two sentences, determine if they have the same meaning.
	question equivalence	Given two questions, determine if they have the same meaning.
NLI - Constraint Satisfaction Problem	house number	There is exactly one person living in each house, and the person living in each house has different characteristics in several dimensions. Given a number of hints (relating the positional relationships of the houses of people with different characteristics), answer the number of the house in which a person with a certain characteristic lives.

Table 2 continued from previous page

Domain	Task Name	Description
NLI - Constraint Satisfaction Problem	logical deduction	Given a paragraph of known conditions (involving the interrelationships of several objects, e.g., location, price, time to accomplish something, age of an antique, etc.), determine which option is correct (each option is a judgment sentence about an object).
LLM Special Test	known unknowns	Factual questions, but some were unknown, testing LLM's ability to answer UNKNOWN.
	hhh alignment	"HHH" stands for 'Helpful, Harmless, and Honest'. Through these tasks, the model can be tested to see if it can be useful, honest, and without negative impacts in real-world applications.
	trolley dilemma	It's a moral question of the Trolley Dilemma type: to do or not to do something.
	color understanding	Given a color representation in RGB/HCL/hexadecimal/HSL format, ask which is the closest color.
Code (Ability to Use)	longest common subsequence	Given two strings, answer the length of the longest common subsequence.
	bracket match judgement	Given a string with parentheses, center brackets, and curly braces, determine if the left and right brackets are perfectly matched.
	bracket match complement	Given a string with parentheses, center brackets, and braces, complete the string so that the left and right brackets match perfectly.
	symbol interpretation	Use different symbols to refer to specific graphics/specific expressions. Given two symbol strings, determine which option's description matches the first string but does not match the second string.

Figure 6: Six examples from the M4CQ dataset demonstrating cross-lingual content consistency. The first and second choices, though semantically similar, are accurately translated, which also reflects the dataset's quality.

Each instance has semantically equivalent versions in all languages in M4CQ, see Figure 6 for examples.

B Comparison of M4CQ and All Existing Multilingual Multitask Datasets

Table 3 details the comparison of M4CQ and all existing multilingual multitask datasets.

C Dataset Format of M4CQ

All tasks are converted to a unified Parquet format with the following schema:

- **`idx`** (Int32): Unique instance identifier within each task.
- **`question`** (String): Task prompt.
- **`choices`** (List[String]): Options, supporting variable counts.
- **`answer`** (List[Int32]): Binary vector indicating correct options (1 for correct, 0 for incorrect), with the same length as options.

This schema supports flexible option counts and future extension to multi-answer questions, overcoming the rigidity of fixed-option formats (e.g., "A/B/C/D" constraints).

D Specifications of the Pre-experiment on Translation Engine Selection

To select suitable translation engines for dataset construction, we conducted a preliminary evaluation across seven translation systems: DeepL, Google Translation, Bing Translation, Claude 3.5 Sonnet, Qwen2.5-72B-Instruct, DeepSeek-R1, and GPT-4o. The experiment involved 1,350 instances which were randomly selected from the M4CQ dataset, which were then translated into five languages: Chinese, Japanese, Turkish, Russian, and Latvian. Each engine produced 6,750 translation outputs (1,350 instances × 5 languages), totaling

Table 3: Comparison of M4CQ and existing multilingual multitask datasets.

Dataset	Languages	Tasks	Instances per Language	Evaluation Modality
MMMLU (OpenAI, 2024)	14	57	≈ 14k	Multiple-choice (4 choices)
Global MMLU (Singh et al., 2024)	42	57	≈ 14k	Multiple-choice (4 choices)
MMLU-ProX (Xuan et al., 2025)	29	14	≈ 12k	Multiple-choice (10 choices)
P-MMEVAL (Zhang et al., 2024)	10	64	3038	Multiple-choice (4 choices) & Text Generation
M4CQ (this work)	**19**	**119**	**≈ 56k**	**Multiple-choice (2∼108 choices)**

Table 4: Translation Engine Performance Comparison (Total Evaluation Number = 6,750 per Engine)

Engine	Perfect (%)	Acceptable with Minor Issues (%)	Erroneous (%)
DeepL	97.35 (6,571)	2.58 (174)	0.07 (5)
GPT-4o	95.96 (6,477)	1.07 (72)	2.98 (201)
DeepSeek-R1	94.79 (6,398)	3.66 (247)	1.56 (105)
Claude 3.5 Sonnet	93.73 (6,327)	5.64 (381)	0.62 (42)
Qwen2.5-72B-Instruct	89.54 (6,044)	6.06 (409)	4.40 (297)
Bing	34.83 (2,351)	28.40 (1,917)	36.77 (2,482)
Google	30.86 (2,083)	33.97 (2,293)	35.17 (2,374)

47,250 translations for human evaluation.

A three-tiered classification method was employed to evaluate the translations:

- **Perfect**: Flawless translation with accurate terminology
- **Acceptable with Minor Issues**: Generally acceptable but containing minor flaws (awkward phrasing or non-standard terminology)
- **Erroneous**: Meaning-altering errors or incomprehensible output

As shown in Table 4, DeepL demonstrated superior performance with 97.35% perfect translations and a remarkably low error rate of 0.07%. GPT-4o, DeepSeek-R1, Claude 3.5 Sonnet, and Qwen2.5-72B-Instruct showed competitive accuracy ranging from 89.54% to 95.96%, though with slightly higher error rates (0.62%–4.40%). Traditional translation services (Bing and Google) exhibited significantly poorer performance, with error rates exceeding 35%.

Based on these findings, DeepL was selected as the primary translation engine due to its exceptional accuracy (high perfect rate) and stability (low error rate). The four language models (GPT-4o, DeepSeek-R1, Claude 3.5 Sonnet, and Qwen2.5-72B-Instruct) were retained for collaborative error correction during quality control phases. Conventional translation engines (Google and Bing) were excluded from subsequent pipeline stages due to their substantially higher error rates.

E Justifications for Excluded Model Families

Some model families are popular but excluded. Here are reasons.

- Llama 3.3: Only contains one model.
- Llama 3.2: The difference in the amount of parameters between the largest and smallest language models within this family does not exceed 10B.
- Qwen 2: Not the most recent iteration in its model family.
- GLM 4.7: Only contains one model.
- GLM 4: Only contains one model.

F Detailed Experimental Results

Table 5 and Table 6 present average accuracy of different models across languages on all tasks.

Table 7 and Table 8 present average accuracy of different models across languages on knowledge-based tasks.

Table 5: Average accuracy of different models across languages on all tasks. (Part 1)

Language	Aya-Expanse-32b	Qwen3-32B	Mistral-Small	Aya-Expanse-8b	Qwen3-14B	Mistral-Nemo
AR	0.480	0.437	0.464	0.431	0.425	0.401
BG	0.419	0.460	0.484	0.364	0.426	0.430
CS	0.490	0.444	0.494	0.441	0.418	0.415
EL	0.490	0.445	0.496	0.431	0.414	0.408
EN	0.551	0.536	0.571	0.499	0.528	0.529
FI	0.398	0.436	0.482	0.341	0.409	0.396
HU	0.402	0.444	0.452	0.335	0.395	0.404
ID	0.484	0.489	0.493	0.450	0.458	0.426
IT	0.505	0.476	0.536	0.454	0.443	0.456
JA	0.464	0.412	0.467	0.418	0.411	0.409
LT	0.391	0.449	0.448	0.336	0.419	0.393
LV	0.378	0.439	0.451	0.329	0.405	0.390
NL	0.495	0.467	0.511	0.446	0.430	0.431
RO	0.512	0.459	0.513	0.451	0.437	0.430
RU	0.488	0.492	0.506	0.432	0.450	0.447
SK	0.463	0.453	0.469	0.399	0.415	0.411
TR	0.466	0.431	0.448	0.417	0.407	0.390
UK	0.483	0.476	0.496	0.437	0.424	0.437
ZH	0.478	0.476	0.494	0.432	0.455	0.442

Table 6: Average accuracy of different models across languages on all tasks. (Part 2)

Language	Phi-4-mini-instruct	Qwen3-1.7B	Qwen3-8B	Qwen3-4B	Qwen3-0.6B	Phi-4
AR	0.386	0.344	0.396	0.389	0.322	0.426
BG	0.387	0.340	0.390	0.393	0.328	0.436
CS	0.393	0.343	0.389	0.384	0.326	0.437
EL	0.357	0.324	0.380	0.363	0.317	0.396
EN	0.527	0.406	0.478	0.489	0.367	0.555
FI	0.392	0.335	0.379	0.374	0.323	0.415
HU	0.395	0.333	0.377	0.367	0.319	0.400
ID	0.398	0.356	0.416	0.402	0.327	0.431
IT	0.419	0.355	0.416	0.395	0.329	0.504
JA	0.398	0.340	0.378	0.378	0.330	0.447
LT	0.354	0.329	0.384	0.377	0.319	0.383
LV	0.343	0.338	0.394	0.373	0.321	0.395
NL	0.429	0.355	0.391	0.389	0.325	0.450
RO	0.393	0.347	0.397	0.399	0.324	0.418
RU	0.427	0.348	0.411	0.401	0.330	0.453
SK	0.379	0.342	0.387	0.373	0.322	0.422
TR	0.381	0.336	0.371	0.368	0.316	0.428
UK	0.395	0.341	0.394	0.377	0.322	0.433
ZH	0.413	0.375	0.435	0.428	0.343	0.451

Table 7: Average accuracy of different models across languages on knowledge-based tasks. (Part 1)

Language	Aya-Expanse-32b	Qwen3-32B	Mistral-Small	Aya-Expanse-8b	Qwen3-14B	Mistral-Nemo
AR	0.420	0.382	0.412	0.369	0.372	0.342
BG	0.369	0.404	0.449	0.300	0.383	0.380
CS	0.436	0.398	0.452	0.379	0.377	0.381
EL	0.432	0.406	0.449	0.378	0.376	0.356
EN	0.505	0.504	0.539	0.447	0.483	0.488
FI	0.329	0.376	0.429	0.273	0.354	0.342
HU	0.337	0.381	0.401	0.278	0.347	0.346
ID	0.433	0.446	0.443	0.388	0.415	0.379
IT	0.462	0.437	0.499	0.407	0.414	0.414
JA	0.410	0.362	0.430	0.358	0.359	0.349
LT	0.323	0.385	0.396	0.278	0.362	0.337
LV	0.312	0.380	0.400	0.271	0.345	0.336
NL	0.448	0.420	0.469	0.392	0.389	0.383
RO	0.465	0.416	0.478	0.404	0.394	0.383
RU	0.442	0.446	0.473	0.377	0.403	0.401
SK	0.406	0.404	0.432	0.339	0.366	0.373
TR	0.413	0.395	0.410	0.364	0.360	0.344
UK	0.432	0.435	0.466	0.377	0.389	0.390
ZH	0.428	0.432	0.457	0.371	0.422	0.385

Table 8: Average accuracy of different models across languages on knowledge-based tasks. (Part 2)

Language	Phi-4-mini-instruct	Qwen3-1.7B	Qwen3-8B	Qwen3-4B	Qwen3-0.6B	Phi-4
AR	0.322	0.286	0.338	0.321	0.263	0.375
BG	0.330	0.283	0.343	0.343	0.267	0.388
CS	0.331	0.292	0.344	0.336	0.262	0.376
EL	0.302	0.269	0.332	0.317	0.259	0.344
EN	0.491	0.363	0.444	0.439	0.309	0.520
FI	0.332	0.276	0.317	0.308	0.261	0.362
HU	0.324	0.278	0.320	0.309	0.256	0.363
ID	0.344	0.300	0.366	0.349	0.268	0.400
IT	0.364	0.309	0.366	0.349	0.277	0.480
JA	0.341	0.291	0.330	0.328	0.276	0.405
LT	0.292	0.279	0.325	0.317	0.260	0.317
LV	0.285	0.275	0.336	0.315	0.260	0.339
NL	0.372	0.302	0.341	0.328	0.268	0.413
RO	0.336	0.293	0.346	0.345	0.262	0.371
RU	0.370	0.297	0.356	0.343	0.274	0.413
SK	0.319	0.283	0.341	0.315	0.264	0.367
TR	0.327	0.283	0.321	0.320	0.264	0.390
UK	0.337	0.290	0.348	0.330	0.269	0.385
ZH	0.352	0.331	0.400	0.390	0.294	0.405