

# DataSmith: Synthesizing Multilingual Instruction-Tuning Data with the Help of External Tools

Lester James V. Miranda<sup>1</sup>, Anna Korhonen<sup>1</sup>

<sup>1</sup>Language Technology Lab, University of Cambridge

Correspondence: ljvm2@cam.ac.uk

## Abstract

Synthesizing supervised finetuning (SFT) data for training multilingual language models (LMs) is common, but this approach relies heavily on the parametric knowledge of the generator model. This reliance is especially limiting for less-resourced languages, where such knowledge is often sparse or unreliable, risking low-quality data and underperforming LMs. In this work, we introduce DATASMITH, a multilingual data generation framework that reduces this reliance by grounding a tool-use teacher LM with external sources. We equip this LM with tools that can access the web, reference linguistic knowledge, and consult other expert systems. Our experiments across 11 languages (5 high resource, 6 low- to mid-resource) show that DATASMITH is capable of producing quality data which can then be used to train multilingual LMs that perform well on downstream cultural knowledge, commonsense reasoning, and knowledge-based tasks. Further analysis of the teacher LM’s trajectory reveals that DATASMITH invokes more tool-use calls and leans more on linguistic resources for less-resourced languages. Ultimately, we demonstrate through the DATASMITH framework that moving away from a model-only paradigm, coupled with a thoughtful curation of external tools, can lead to quality synthetic data for building the next generation of inclusive language technologies.

Status: In Progress

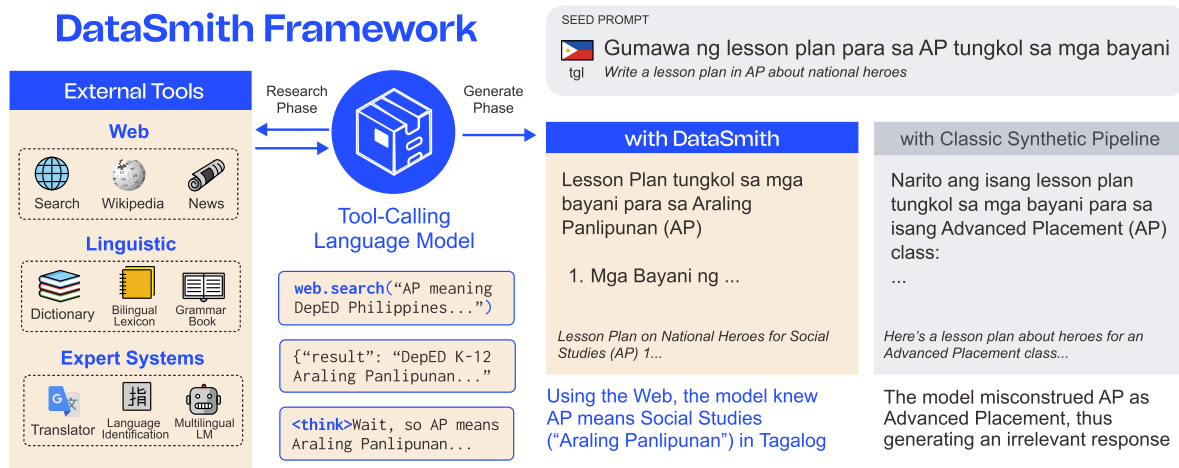
Website: <https://ljvmiranda921.github.io/datasmith>

GitHub: <https://github.com/ljvmiranda921/datasmith>

HuggingFace: <https://huggingface.co/ljvmiranda921/datasmith>

## 1 Introduction

Synthetic data generation has become a dominant paradigm for generating cost-efficient datasets for adapting language models (LMs) across several languages (Martins et al., 2025; Chitale et al., 2026, *inter alia*). This approach typically involves prompting a **teacher LM** to produce simulated, multi-turn conversations between a user and an assistant. The generated dataset is then used to adapt a base model using methods such as supervised finetuning (SFT, Ouyang et al., 2022). However, relying solely on the parametric knowledge of the teacher has several limitations: a model not fluent in the language or its cultural dimensions risks producing texts that are unnatural or lacking cultural nuance (Liu et al., 2025; Pawar et al., 2025). This limitation becomes more apparent in the case of low-resource languages, where pretraining representation is already small and noisy (Kreutzer et al., 2022), resulting in low-quality data and underperforming LMs (Miranda et al., 2026). Several works have tried to address this problem by optimizing various parts of the synthetic data pipeline such as the construction of prompts (Mora et al., 2025), the choice of teacher LM (Cegin et al., 2026; Miranda et al., 2026), or the data-filtering critic (Pombal et al., 2025; Anugraha et al., 2026). On the other hand, frontier LMs have excelled in **tool-use**, i.e., tasks that involve interacting with and reasoning about external applications to solve a task. For example, frontier LMs can interact with web APIs and real-world applications across many domains (Trivedi et al., 2024; Patil et al., 2025; Yao et al., 2025). While prior works have explored strategies to integrate tools for data generation or annotation (Mitra et al., 2024; Findeis et al., 2025, *inter alia*), none have done so in the multilingual context, where harnessing external knowledge beyond the teacher’s capabilities is crucial.



**Figure 1 Overview of the DataSmith framework and its comparison to a standard synthetic data pipeline.** While a classic synthetic pipeline is restricted to its own parametric knowledge when generating data, DATASMITH is augmented with external tools such as web search, linguistic resources, and expert systems, grounding the generation process in external knowledge.

In this work, we introduce **DataSmith** (Figure 1), a framework for **S**ynthesizing **M**ultilingual **I**nstruction-**T**uning with the **H**elp of **E**xternal **T**ools. Specifically, we equip a tool-use LM with functions to access the **Web** (Wikipedia, Web Search, News Search), refer to **Linguistic Resources** (Bilingual Lexicons, Dictionary, Grammar Books), and consult **Expert Systems** (Translation Model, Language ID, Larger Multilingual Model). Using this framework, we synthesized multi-turn dialogues for SFT on 11 languages (5 high-resource, 6 low- to mid-resource, according to Joshi et al., 2020), training instruction-tuned models that outperform, on average, a classic synthetic data pipeline across knowledge-based (Global-MMLU Lite, Singh et al., 2025), commonsense reasoning (Global PIQA, Chang et al., 2025), and cultural benchmarks (CulturalBench, Chiu et al., 2025).

Our contributions are as follows:

- We introduce a **synthetic data framework** (DATASMITH, §4) that integrates tool-use and apply it in the context of multilinguality. To the best of our knowledge, this is the first work that harnesses LM externalities to generate data for low- to medium-resource languages. In downstream evaluations (§5), the datasets produced by DATASMITH outperform those from a classic synthetic data generation pipeline driven by a strong teacher LM (Miranda et al., 2026), as well as a standard translate-train setup.
- We provide **analyses and insights** into the tool-use behavior of LMs when asked to generate multilingual data (§6). For example, we find that DATASMITH adapts its tool-use to the target language, issuing more calls and leaning more on linguistic resources for low- to mid-resource languages than for high-resource ones (§6.1). We also find that the resulting research trajectories are well-grounded in retrieved evidence across all languages (§6.2).
- We also **close the resource gap** by releasing SFT datasets for 11 languages to stimulate the development of more equitable LMs. We also include DATASMITH trajectories to enable further research in agent-based data generation.

We plan to open-source our framework and release the synthetic datasets in order to pave the way for inclusive and equitable language technologies through the creation of high-quality data.

## 2 Background: Tools for Constructing Multilingual Data

Constructing multilingual datasets typically draws on a range of sources. Building on prior surveys (Yu et al., 2022; Wang et al., 2025; Pava et al., 2026), we group these into three broad categories: the web, linguistic resources, and expert systems. We provide a summary of prior work in Table 1.

**Table 1 External tools used in multilingual dataset construction.** Based on the taxonomy from prior works (Yu et al., 2022; Wang et al., 2025; Pava et al., 2026), we organize the tools across three categories and show representative work that constructed datasets by harnessing each tool.

| Category       | Tool            | Prior Works                                                                   |
|----------------|-----------------|-------------------------------------------------------------------------------|
| Web            | Search          | Artetxe et al. (2022); Li et al. (2023)                                       |
|                | Wikipedia       | Pan et al. (2017); Schwenk et al. (2021); Zhang et al. (2023)                 |
|                | News            | Hasan et al. (2021); Hwang et al. (2024); Bozkurt (2024)                      |
| Linguistic     | Dictionary      | Zhang et al. (2024); Ghazvininejad et al. (2023)                              |
|                | Lexicon         | Lu et al. (2024); Wang et al. (2022); Jones et al. (2023); Yong et al. (2024) |
|                | Grammar Book    | Hus and Anastasopoulos (2024); Tanzer et al. (2024)                           |
| Expert Systems | Translator      | Artetxe et al. (2023); Muennighoff et al. (2023); Lai et al. (2023)           |
|                | Language ID     | Caswell et al. (2020); Burchell et al. (2023); Kargaran et al. (2023)         |
|                | Multilingual LM | Chitale et al. (2026); Miranda et al. (2026)                                  |

**Web** is the most common venue for sourcing text in many languages. General web search has been used to assemble higher-quality corpora for low-resource languages through tailored crawling (Artetxe et al., 2022) and seeded discovery of educational material (Li et al., 2023). Wikipedia is a particularly rich source given its broad multilingual coverage, and has been mined for several applications such as cross-lingual name tagging (Pan et al., 2017), parallel sentence construction (Schwenk et al., 2021), and text retrieval (Zhang et al., 2023). News articles are another source, supporting the construction of text summarization datasets (Hasan et al., 2021) and fact-checking QA data (Hwang et al., 2024; Bozkurt, 2024).

**Linguistic resources** encode knowledge about a language, and are valuable for low-resource and endangered languages. Dictionaries have been used to prompt LLMs for translation at the word level (Ghazvininejad et al., 2023; Lu et al., 2024) among other things. Bilingual lexicons offer broader language coverage than parallel text, and have been used to extend pretrained models to thousands of new languages (Wang et al., 2022; Jones et al., 2023; Yong et al., 2024). Grammar books are the most structured, and recent work has shown that long-context LMs can leverage them to translate languages absent from pretraining (Tanzer et al., 2024; Hus and Anastasopoulos, 2024).

**Expert systems** are trained models that perform specific multilingual functions. Machine translation systems are commonly used to translate English instruction data into target languages (Muennighoff et al., 2023; Lai et al., 2023). Language ID models filter and label multilingual corpora at scale, with recent systems extending coverage past 1,000 languages (Caswell et al., 2020; Kargaran et al., 2023; Burchell et al., 2023). Finally, larger multilingual LMs themselves serve as expert systems for data generation, supplying the seed prompts and responses that smaller student models are then trained on (Chitale et al., 2026; Miranda et al., 2026).

### 3 Related Work

**Multilingual LM Adaptation and the Role of Synthetic Data** A common approach for adapting LMs to another language is to take a pretrained base model ( $S_\phi$ ) and perform supervised finetuning (SFT, Ouyang et al., 2022) on an instruction-tuning dataset. This dataset contains prompt-response pairs  $(x, y)$ , where the prompt  $x$  is a user request or instruction and the response  $y$  is the LM’s generated answer. For multilingual tasks, such datasets can be curated by native speakers (Singh et al., 2024) or produced via synthetic data generation (Cahyawijaya et al., 2024; Chitale et al., 2026, *inter alia*). The latter is often seen as an appealing alternative to human annotation given its low cost and scalability. Miranda et al. (2026) provide a brief survey of common multilingual synthetic data pipelines, including generating new instances from demonstrations, translating from English, and responding to a prompt in the target language.

**Grounding Synthetic Data Generation with External Knowledge** Several works have attempted to ground the synthetic data pipeline in external knowledge such as unstructured documents (Lupidi et al., 2025; Yuan

et al., 2026) or terminal and calculator outputs (Findeis et al., 2025), in order to reduce hallucination and improve the diversity of generated examples. Other recent approaches employ *agentic data pipelines* (Luo et al., 2024; Lù and Reddy, 2026; Kulikov et al., 2026), which combine reasoning, tool-use, memory management, and long-horizon interactions to generate synthetic data for coding (Mitra et al., 2024) and scientific tasks (Kulikov et al., 2026). Under this paradigm, data generation is no longer solely a function of the teacher LM’s parametric knowledge, as it draws on external information assumed to be grounded.

## 4 The DataSmith Framework

**System Definition** Building on the agent formalism introduced in prior works (Guo et al., 2024; Kim et al., 2026; Zhu et al., 2026), we define DATASMITH as a single-agent system tuple  $S = (\Phi, M, \mathcal{A}, \pi)$ , where:

- $\Phi$  is the reasoning policy (typically a large language model)
- $M$  is the internal memory, in this case, the context length of  $\Phi$ .
- $\mathcal{A} = \{\text{ToolCall}(t, \theta) \in \mathcal{T}, \theta \in \Theta_t\}$  is the action space consisting of tool-use, with  $\mathcal{T}$  as the set of available tools (§4.1) and  $\Theta_t$  represents the parameter configurations for  $t$  (Table 10).
- $\pi : \mathcal{H} \rightarrow \mathcal{A}$  is a decision function mapping histories of past actions and observations to tool-use actions.

Given a language model with a reasoning policy  $\Phi$  and context  $M$ , the DATASMITH framework introduces two components: a (1) set of external tools ( $\mathcal{T}$ , §4.1) that can be used for verifying, querying, and retrieving information, and a (2) generation pipeline ( $\mathcal{A}, \pi$ , §4.2) that enables a teacher LM to harness these tools and synthesize multi-turn conversations.

### 4.1 External Tools

We partition the set of available tools into three categories. Detailed information, including API arguments, can be found as supplementary in Table 10.

**Web:** this set of tools provides the teacher LM with access to information from the Web. For general web (Web Search) and news (News Search) queries, we use the DDGS library<sup>1</sup>, which routes searches through proxy backends such as Brave and DuckDuckGo. To fetch pages from Wikipedia (Wikipedia), we use the Wikipedia API, restricting outputs to the English edition because it offers broader coverage and is generally considered more reliable than smaller-language editions (Roy et al., 2022; D’Ignazi et al., 2026).

**Linguistic Resources:** this set of tools enables the teacher LM to verify usage and improve its fluency when generating conversations in languages other than English. To obtain lexical definitions in the target language, we provide a dictionary tool (Dictionary) backed by the Wiktionary API. We also construct bilingual pairs (Bilingual Lexicon) mapping English to the target language using the PanLex dataset (Kamholz et al., 2014). Finally, we incorporate a collection of grammar books and instructional materials released under permissive licenses (Table 9), which we parse with PyMuPDF and retrieve over using BM25.

**Expert Systems:** a function-calling teacher LM may lack the ability to handle multilingual data generation on its own. To address this limitation, we expose a set of specialized expert systems, typically other models that the generator can query. For language identification (Language ID), we use the GlotLID model (Kargaran et al., 2023), which offers broad coverage of low-resource languages. For translation (Translate), we provide access to an expert model, TranslateGemma 27B (Finkelstein et al., 2026). Finally, we allow the teacher to consult a separate LM (Multilingual LM), Qwen 3 32B (Yang et al., 2025), for additional support.

<sup>1</sup><https://pypi.org/project/ddgs/>

**Table 2 Sources of English seed prompts used in DataSmith.** Breakdown of the English instruction-tuning datasets from which seed prompts are drawn, along with the number of prompts sampled from each source.

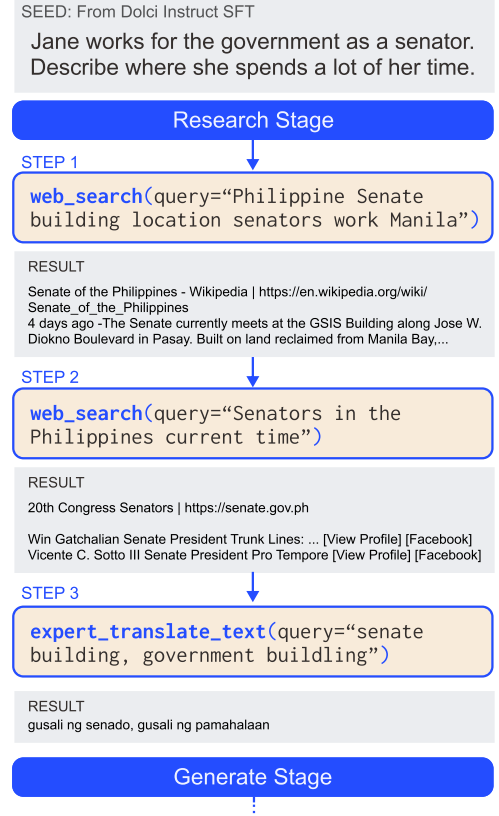
| Source                                                         | # Prompts        | License               |
|----------------------------------------------------------------|------------------|-----------------------|
| Nemotron Instruction Following Chat (NVIDIA et al., 2025)      | 426,009          | CC-BY 4.0, ODC-BY 1.0 |
| Wildchat (Zhao et al., 2024)                                   | 285,139          | ODC-BY 1.0            |
| Dolci Instruct OpenThoughts3+ Science (Team Olmo et al., 2026) | 96,006           | ODC-BY 1.0            |
| FLAN (Wei et al., 2022)                                        | 84,968           | Apache 2.0            |
| OpenMathInstruct 2 (Toshniwal et al., 2025)                    | 50,000           | CC-BY-4.0             |
| Tulu 3 Persona MATH (Lambert et al., 2025)                     | 30,000           | ODC-BY 1.0            |
| Dolci Logic Puzzles (Team Olmo et al., 2026)                   | 20,000           | ODC-BY 1.0            |
| Tulu 3 Persona GSM (Lambert et al., 2025)                      | 20,000           | ODC-BY 1.0            |
| Tulu 3 Persona Algebra (Lambert et al., 2025)                  | 19,994           | ODC-BY 1.0            |
| TableGPT (Li et al., 2024)                                     | 4,580            | MIT                   |
| SciRiff (Wadden et al., 2025)                                  | 4,273            | ODC-BY 1.0            |
| Aya (Singh et al., 2024)                                       | 2,504            | Apache 2.0            |
| OpenAssistant (Köpf et al., 2023)                              | 2,223            | Apache 2.0            |
| <b>Total</b>                                                   | <b>1,045,696</b> |                       |

## 4.2 Generation Pipeline

**Seed Dataset** Each episode begins with an English seed prompt  $x \sim \mathcal{D}_{\text{seed}}$ , which initializes the history  $h_0 = x$ . We draw seed prompts from open datasets curated for LM post-training, namely the Dolci dataset from OLMo 3 (Team Olmo et al., 2026) and the NVIDIA Nemotron datasets (NVIDIA et al., 2025); Table 2 lists all sources. Because many of these prompts are tied to English-language or culturally specific contexts that do not transfer cleanly to other languages, we treat them as inspiration for the teacher LM rather than as strict templates to be translated.

**Data Generation** Given this seed dataset, DATASMITH’s generation pipeline has two stages. In the **Research Stage**, at each step  $t$ , the teacher LM selects a tool-use action  $a_t = \pi(h_t)$  and appends the resulting observation  $o_t$  to form  $h_{t+1}$ . The phase terminates at step  $T$  when the LM signals that sufficient context has been gathered, or when a time budget elapses (Figure 2). In the **Generate Stage**, guided by a system prompt (Figure 6) and a generation prompt (Figure 7),  $\Phi$  produces a multi-turn conversation  $y$  in the target language conditioned on  $h_T$ . We then apply a post-hoc filter  $f(y)$  to discard conversations that depart from the ChatML format, as well as those prompts not in the target language.

This pipeline yields multi-turn conversations  $y$  spanning a diverse set of topics in the target language, each paired with the teacher LM’s research-phase trajectory  $\tau = h_T$ . We will publicly release both the conversations and their trajectories to support future work on agentic data generation.



**Figure 2 An example of the Research Stage.** Given a seed prompt, DATASMITH thinks and plans, issues a sequence of tool-use calls, then proceeds to the Generate Stage where it synthesizes the conversation in the target language.

**Table 3 Main results across all languages.** Accuracy of  $S_\phi$  = Gemma 3 12B Instruct fine-tuned on data from each method, averaged across Global MMLU, Global PIQA, and CulturalBench and over three runs. *Avg* is the macro-average across all languages.

| Method                       | Avg         | High        |             |             |             |             | Low-Mid     |             |             |             |             |             |
|------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                              |             | ara         | deu         | fra         | spa         | jpn         | ces         | ind         | tgl         | ceb         | swa         | yor         |
| $S_\phi$ (no SFT)            | 63.2        | 66.1        | 67.9        | 64.0        | 72.0        | 71.2        | 57.7        | 70.9        | 55.4        | 45.6        | 71.3        | 53.4        |
| + DATASMITH (Ours)           | <b>67.9</b> | <b>68.3</b> | <b>70.2</b> | 70.3        | <b>74.6</b> | <b>73.8</b> | <b>66.0</b> | <b>74.8</b> | <b>64.7</b> | <b>52.2</b> | <b>72.3</b> | <b>60.2</b> |
| + Teacher-only $\Phi_\theta$ | 65.4        | 67.7        | 68.3        | <b>70.8</b> | 71.8        | 72.7        | 64.1        | 72.5        | 57.7        | 48.9        | 68.2        | 56.8        |
| + Translate-Train            | 55.5        | 53.2        | 62.3        | 63.6        | 62.7        | 60.1        | 55.0        | 67.0        | 55.9        | 31.1        | 62.0        | 37.4        |

**Table 4 Results for each benchmark, broken down by resource-level.** Downstream performance of  $S_\phi$  = Gemma 3 12B Instruct fine-tuned on data from each method, averaged over three runs per language. *Avg* is the macro-average across resource-levels. We show standard errors and breakdowns for each language in Table 11.

| Method                       | Global MMLU |             |             | Global PIQA |             |             | CulturalBench |             |             |
|------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|---------------|-------------|-------------|
|                              | High        | Low-Mid     | Avg         | High        | Low-Mid     | Avg         | High          | Low-Mid     | Avg         |
| $S_\phi$ (no SFT)            | 60.7        | 45.5        | 53.1        | 81.6        | 79.2        | 80.4        | 61.5          | 55.5        | 58.5        |
| + DATASMITH (Ours)           | <b>65.3</b> | <b>56.9</b> | <b>61.1</b> | <b>82.8</b> | <b>82.4</b> | <b>82.6</b> | <b>65.7</b>   | <b>59.5</b> | <b>62.6</b> |
| + Teacher-only $\Phi_\theta$ | <b>65.3</b> | 52.9        | 59.1        | 80.7        | 77.6        | 79.2        | 64.0          | 57.2        | 60.6        |
| + Translate-Train            | 57.7        | 53.7        | 55.7        | 73.8        | 68.6        | 71.2        | 48.8          | 37.9        | 43.4        |

## 5 Experiments: Does DataSmith Generate Better Supervision Data?

In order to test whether DATASMITH is a better alternative for multilingual synthetic data generation, we measure dataset quality, which we ground to the performance of a *student* model trained from this dataset on downstream tasks. We test on 11 languages (5 high-resource and 6 low-mid resource): Arabic, German, French, Spanish, Japanese, Czech, Indonesian, Tagalog, Cebuano, Swahili, and Yoruba.

**Pipeline Setup** We use Kimi-K2.5 as the main tool-use model ( $\Phi$ ). We generate 5k multi-turn dialogues for each language. This choice is based on works examining the effect of dataset size to elicit multilingual performance (Shaham et al., 2024; Shimabucoro et al., 2025). For our main experiments, we perform continuous SFT on Gemma 3 12B Instruct as our base model ( $S_\phi$ ).

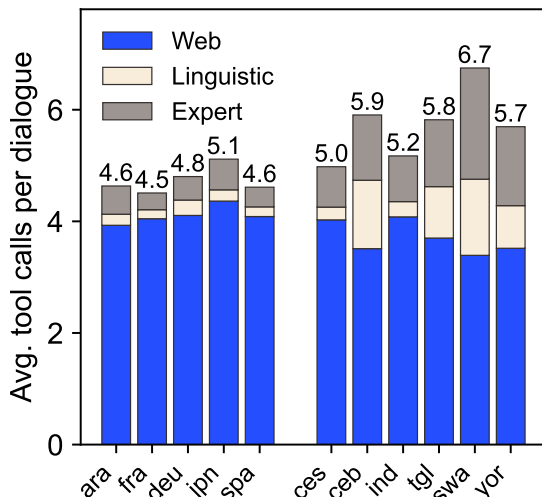
**Benchmarks** We evaluate on the following downstream tasks: factual knowledge (Global MMLU, Singh et al., 2025), physical reasoning (Global PIQA, Chang et al., 2025), and cultural knowledge (CulturalBench, Chiu et al., 2025). We choose these benchmarks because they cover the majority of the languages in our experiment. We report the average score per language across three runs.

**Baselines** We compare DATASMITH against two alternative ways of producing SFT data, training a separate student model on each:

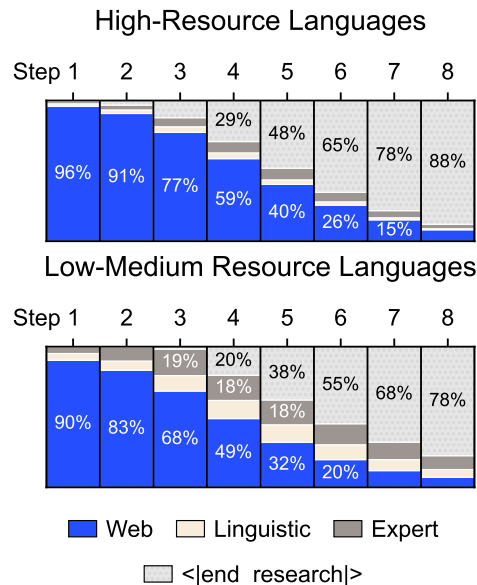
- *Teacher-only  $\Phi_\theta$* : the same teacher LM ( $\Phi$ ), but without tool access, prompted to generate 5k conversations per language (see Figure 8 for the teacher-only prompt).
- *Translate-Train*: we translate 5k conversations from the English seed prompts  $\mathcal{D}_{\text{seed}}$  into each target language using TranslateGemma 27B (Finkelstein et al., 2026).

All baselines use the same base model, Gemma 3 12B Instruct, and hyperparameters to ensure a fair comparison. We additionally report the base model performance as reference. Appendix C contains more information in our experimental and evaluation setup.

**Results** DATASMITH improves over the base model and outperforms all alternative synthetic data baselines, both across languages and across benchmarks. It achieves the best score in 10 of 11 languages (Table 3),



**Figure 3 Tool use by language.** Average number of tool-use calls per dialogue, broken down by tool category, for high-resource (*Left*) and low- to mid-resource (*Right*) languages. Low- to mid-resource languages invoke more tools overall and rely more on linguistic resources, while web tools dominate across all languages.



**Figure 4 Tool use across the research stage.** The teacher LM’s action at each step of the research stage, with `<|endturn|>` marking turns where it is satisfied and ends research. The teacher leans heavily on web search in early turns. When generating for low- to mid-resource languages, it researches for longer and relies more on linguistic and expert tools.

trailing only on French ( $-0.5\text{pp}$ ). **The largest gains come from low- to mid-resource languages** such as Tagalog ( $+9.3\text{pp}$ ), Czech ( $+8.3\text{pp}$ ), Yoruba ( $+6.8\text{pp}$ ), and Cebuano ( $+6.6\text{pp}$ ). DATASMITH also **ranks first on all three benchmarks** (Table 4), with the largest improvement on Global MMLU ( $45.5 \rightarrow 56.9$ ,  $+11.4\text{pp}$ ). Interestingly, we also notice that *Translate-Train* models struggle with instruction-following on all benchmarks during evaluation, hence, we implement a best-effort parser for their outputs. Furthermore, we also show that our findings generalize on other base model families such as Tiny Aya Global 3B (Table 12, Salamanca et al., 2026) and Llama 3.1 (Table 13, Grattafiori et al., 2024). Overall, our findings suggest that grounding synthetic data generation in tool-use yields better supervision data than raw teacher knowledge or translation alone, with the clearest benefits for underserved languages.

## 6 Analysis of Research Trajectories

### 6.1 How does tool-use vary across languages and research steps?

**Setup** In order to understand DATASMITH’s tool-use patterns, we track its *research trajectory*, i.e., a sequence of tool-use calls it issues before writing each dialogue. We assign every call to one of three families (Table 10), and for each step we compute the share of dialogues that invoke each family or that have already stopped (`<|end_research|>`), averaged across the languages in each resource tier.

**Results** Figure 3 shows the average number of tool-use calls per dialogue for high- and low- to medium-resource languages. We find that DATASMITH **invokes more tool-use calls for low- to medium-resource languages** (5.0–6.7 on average) than for high-resource ones (4.5–5.1), and that a larger fraction of these calls go to linguistic resources. This suggests that low- to mid-resource languages require grounding in the language itself (e.g., vocabulary and usage), rather than in factual knowledge alone. When looking at each step in Figure 4, Web dominates the first few turns across all languages, but there is **more diversity in the tool-use calls for low- to medium-resource languages in later steps**. Together, these results show that

DATA SMITH adapts its research to the target language: for low- to medium-resource languages, it issues more tool-use calls and shifts them beyond web search toward linguistic and expert tools as research progresses.

## 6.2 How good is DataSmith’s research?

**Setup** In order to measure the quality of DATA SMITH’s research stage, Inspired by Search-R2 (He et al., 2026), we employ an LLM-as-a-judge that scores each trajectory using the following rubrics (scored from 1–5):

- *Evidence groundedness*: whether the content of the synthesized dialogue are supported by information the teacher actually retrieved through tools rather than asserted from its priors.
- *Information density*: the proportion of the teacher’s retrieved information that is relevant and useful for producing the sample.
- *Research efficiency*: how efficiently the trajectory uses its tool-use calls. Also judges whether redundant tool-use calls were made by the teacher.

We sample 1k trajectories for each language and score them using GPT-OSS 120B (OpenAI et al., 2025). When prompting the judge, we provide few-shot exemplars of reference research trajectories with human-annotated scores in order to guide the model when scoring new trajectories (see Appendix D). We ran the experiment for three trials.

**Measuring Judge Reliability** In order to test whether an LLM judge is enough to determine the quality of the research trajectories, we compare the annotations of a human annotator and GPT-OSS 120B. The overall inter-annotator agreement (IAA) is 0.769 (Weighted Cohen’s  $\kappa$ , full results in Table 5). We find high agreement across all rubrics, suggesting that GPT-OSS 120B is a reliable judge for research trajectory quality.

**Results** Figure 5 shows the average score per criterion over 1k trajectories for each language. In general, we find that the **research trajectories are well-grounded in evidence** across most languages. This suggests that most of the SFT conversations synthesized during the Generate phase rely on information actually gathered during research rather than the teacher’s priors. However, **research efficiency is the weakest of the three rubrics**, and this drop is more pronounced for low-resource languages such as Yoruba and Swahili (below the divider). An inefficient research stage tends to issue redundant tool-use calls, often with similar parameters. We show an example for Yoruba in Figure 11. Overall, we find the quality of DATA SMITH’s research trajectories to be satisfactory, with scores ranging from 3.35 to 3.77.

## 7 Discussion and Future Work

By treating multilingual data generation as tool-augmented research, DATA SMITH opens up a broader design space, most of which we leave to future work. We highlight a few directions that we find especially promising.

**Automated tool-choice and harness design for data generation** Thus far, we have based our tool set on sources commonly used to construct multilingual data (§2). Our analysis in §6.1, however, shows that the

**Table 5 Inter-annotator agreement between a human annotator and GPT-OSS 120B** We report the Quadratic Weighted Cohen’s  $\kappa$  across 100 instances.

| Rubric                 | Wt. Cohen’s $\kappa$ |
|------------------------|----------------------|
| Evidence Groundedness  | 0.743                |
| Information Density    | 0.810                |
| Research Efficiency    | 0.755                |
| <b>Overall Average</b> | <b>0.769</b>         |

|     | Evidence Groundedness | Information Density | Research Efficiency |
|-----|-----------------------|---------------------|---------------------|
| ara | 4.16                  | 3.75                | 3.37                |
| jpn | 4.08                  | 3.72                | 3.21                |
| spa | 4.17                  | 3.81                | 3.40                |
| fra | 4.14                  | 3.75                | 3.41                |
| deu | 4.14                  | 3.76                | 3.33                |
| yor | 3.92                  | 3.32                | 2.81                |
| swa | 3.99                  | 3.45                | 2.78                |
| ceb | 4.15                  | 3.49                | 2.83                |
| tgl | 4.14                  | 3.56                | 2.93                |
| ind | 4.25                  | 3.78                | 3.25                |
| ces | 4.17                  | 3.73                | 3.21                |

**Figure 5 Quality of DataSmith’s research trajectories across languages (min: 1.0, max: 5.0).** We report the mean LLM-as-a-judge score (GPT-OSS 120B) over 1k trajectories per language along three axes.

teacher does not rely on all tools equally: web search dominates across all languages, while linguistic resources are invoked far more for low- to mid-resource languages (Figure 3). This suggests that the optimal set of tools could be discovered automatically, conditioned on the difficulty of a language or the data generation task. For instance, a harness could allocate more linguistic and expert tools to the lower-resource languages that lean on them most, concentrating grounding where our results show it yields the largest gains.

**Training a data generator LM** In this work, we rely on a fixed, prompted generalist LM with tool-use capabilities as our teacher. A promising alternative is to train a task-specific data generator from the research trajectories we collect and the rubrics we develop for assessing their quality (§6.2). Because downstream student performance provides a natural reward signal, such a generator could be optimized with reinforcement learning to produce trajectories that yield better student models in order to close the loop between data generation and downstream model evaluation. Contemporary work by Kulikov et al. (2026) explores a similar idea, which we believe can be extended to multilingual settings.

## 8 Conclusion

In this work, we introduced DATASMITH, a multilingual data generation framework that grounds a teacher LM in external knowledge through tool-use, drawing on the web, linguistic resources, and expert systems. Across 11 languages and three benchmarks, student models trained on DATASMITH-generated data outperform those trained with a classic teacher-only pipeline and a translate-train setup, with the largest gains on low- to mid-resource languages. Our analysis of the teacher’s research trajectories shows that this grounding is adaptive: the teacher issues more tool calls and leans more heavily on linguistic resources specifically for the languages where its parametric knowledge is weakest. Together, these results suggest a path to better multilingual supervision data: equipping teachers with access to the right external sources. We hope DATASMITH paves the way toward more equitable language technologies for underserved languages.

## Limitations

Our work comes with limitations. First, the quality of the data generated by DATASMITH is dependent on the design of its harness and the choice of external tools. We mitigate this by grounding our tool choice in sources commonly used for constructing multilingual datasets (§2), but alternative tool combinations may yield different results, which we leave for future work. Second, our framework assumes that these tools adequately cover the target language. This assumption weakens as we move down the resource hierarchy as Wikipedia coverage varies widely in both size and quality across languages and many languages lack digitized grammar books or lexicons. In addition, expert translation and language identification models only support a subset of the world’s languages. Finally, our experiments use a teacher LM with strong function-calling capabilities over long horizons, such as Kimi-K2.5. Weaker teachers may fail to execute the research pipeline reliably, and we have not characterized how data quality degrades as the teacher’s tool-use ability decreases.

## Ethics Statement

Synthetic data generation carries the risk of perpetuating biases present in the models involved in the pipeline. Although DATASMITH grounds generation in external sources, the teacher LM, the expert systems it consults, and even sources such as web search results and Wikipedia can encode societal and cultural biases, which may propagate into the generated conversations and, ultimately, into the student models trained on them. We partially mitigate this through grounding, but we did not perform a dedicated audit of the generated datasets, and we encourage practitioners to apply appropriate filtering and auditing before deployment.

## Acknowledgments

LJVM and AK acknowledge the support of the UKRI Frontier Grant EP/Y031350/1 (EQUATE). This work was performed using the Isambard AI National AI Research Resource (AIRR) ST/AIRR/I-A-I/1023, and the

Microsoft Research Azure AI Grant. LJVM would also like to thank Jacki O’Neill, Millicent O’Chieng, and Felermino Ali for helpful discussions on the project.

## References

- Agustini, Juliana Wijaya, and Rajiv Ranjan. 2026. [Advanced Indonesian](#). Michigan State University Libraries. ISBN: 978-1-62610-135-7.
- David Anugraha, Shou-Yi Hung, Zilu Tang, En-Shiun Annie Lee, Derry Tanti Wijaya, and Genta Indra Winata. 2026. [mR3: Multilingual Rubric-Agnostic Reward Reasoning Models](#). In *The Fourteenth International Conference on Learning Representations*.
- Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. [Does Corpus Quality Really Matter for Low-Resource Languages?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. [Revisiting Machine Translation for Cross-lingual Classification](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.
- Betty Baura, Ginny Cyboski, David Flaccus, and Lani Keetch. 1967. [Cebuano para sa mga Peace Corps Volunteers \(Cebuano for the Peace Corps Volunteers\)](#). Technical Report ED012905, US Department of Health, Education, and Welfare, Office of Education, Washington, DC.
- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2026. [Lessons from the Trenches on Reproducible Evaluation of Language Models](#). *Preprint*, arXiv:2405.14782.
- Leonard Bloomfield. 1917. [Tagalog Texts with Grammatical Analysis](#), volume 3 of *University of Illinois Studies in Language and Literature*. University of Illinois, Urbana, IL. Public domain. Project Gutenberg eBook #50797.
- Basak Bozkurt. 2024. [A Thesis Proposal ClaimInspector Framework: A Hybrid Approach to Data Annotation using Fact-Checked Claims and LLMs](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 215–224, St. Julian’s, Malta. Association for Computational Linguistics.
- Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. [An Open Dataset and Model for Language Identification](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.
- Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, Dea Annisayanti Putri, Bryan Wilie, Genta Indra Winata, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2024. [Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. [Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Jan Cegin, Branislav Pecher, Ivan Srba, and Jakub Simko. 2026. [RoSE: Round-robin Synthetic Data Evaluation for Selecting LLM Generators without Human Test Sets](#). In *Proceedings of the 19th Conference*

- of the *European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5530–5545, Rabat, Morocco. Association for Computational Linguistics.
- Tyler A. Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanahun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Aftab Iyigun, Afonso Simplicio, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, and 319 others. 2025. [Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures](#). *Preprint*, arXiv:2510.24081.
- Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. 2025. [Beyond the Surface: Measuring Self-Preference in LLM Judgments](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672, Suzhou, China. Association for Computational Linguistics.
- Pranjal A. Chitale, Varun Gumma, Sanchit Ahuja, Prashant Kodali, Manan Uppadhyay, Deepthi Sudharsan, and Sunayana Sitaram. 2026. [UPDESH: Synthesizing Grounded Instruction Tuning Data for 13 Indic Languages](#). *Preprint*, arXiv:2509.21294.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Schwartz, and Yejin Choi. 2025. [CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs’ Cultural Knowledge Through Human-AI Red-Teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.
- Samuel Crowther. 1843. [A Vocabulary of the Yoruba Language](#). Church Missionary Society, London.
- Samuel Crowther. 1852. [A Grammar of the Yoruba Language](#). Seeleys, London.
- Defense Language Institute, Foreign Language Center. 1985a. [DLI Headstart for the Philippines: Cultural Notes](#). Defense Language Institute, Foreign Language Center, Monterey, CA.
- Defense Language Institute, Foreign Language Center. 1985b. [Headstart for the Philippines: Modules I-V](#), first edition. Defense Language Institute, Foreign Language Center, Monterey, CA.
- Jacopo D’Ignazi, Andreas Kaltenbrunner, Yelena Mejova, Michele Tizzani, Kyriaki Kalimeri, Mariano G. Beiró, and Pablo Aragón. 2026. [Language-Agnostic Modeling of Source Reliability on Wikipedia](#). *ACM Trans. Web*, 20(1).
- Arduin Findeis, Floris Weers, Guoli Yin, Ke Ye, Ruoming Pang, and Tom Gunter. 2025. [Can external validation tools improve annotation quality for LLM-as-a-judge?](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15997–16020, Vienna, Austria. Association for Computational Linguistics.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. [TranslateGemma Technical Report](#). *Preprint*, arXiv:2601.09012.
- Foreign Service Institute. 1970. *Indonesian Basic Course: The Djakarta Dialect*. U.S. Department of State, Foreign Service Institute, Washington, DC.
- 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](#).
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. [Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation](#). *Preprint*, arXiv:2302.07856.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
-

- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. [Large Language Model Based Multi-agents: A Survey of Progress and Challenges](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 8048–8057. International Joint Conferences on Artificial Intelligence Organization. Survey Track.
- Daniel Han, Michael Han, and Unsloth Team. 2023. [Unsloth](#).
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. [XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.
- Bowei He, Minda Hu, Zenan Xu, Hongru Wang, Licheng Zong, Yankai Chen, Chen Ma, Xue Liu, Pluto Zhou, and Irwin King. 2026. [Search-R2: Enhancing Search-Integrated Reasoning via Actor-Refiner Collaboration](#). *Preprint*, arXiv:2602.03647.
- Christian Hilchey, Petra Mert, Tracy Heim, Ryan Williams, Carl Blyth, Sarah Sweeney, and Nathalie Steinfeld Childre. 2018. [Reality Czech: An Open Czech Curriculum](#). Online textbook.
- Jonathan Hus and Antonios Anastasopoulos. 2024. [Back to School: Translation Using Grammar Books](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.
- Alyssa Hwang, Kalpit Dixit, Miguel Ballesteros, Yassine Benajiba, Vittorio Castelli, Markus Dreyer, Mohit Bansal, and Kathleen McKeown. 2024. [NewsQs: Multi-Source Question Generation for the Inquiring Mind](#). *Preprint*, arXiv:2402.18479.
- Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. [GATITOS: Using a New Multilingual Lexicon for Low-resource Machine Translation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The State and Fate of Linguistic Diversity and Inclusion in the NLP World](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- David Kamholz, Jonathan Pool, and Susan Colowick. 2014. [PanLex: Building a Resource for Panlingual Lexical Translation](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. [GlotLID: Language Identification for Low-Resource Languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.
- Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. 2026. [Towards a Science of Scaling Agent Systems](#). *Preprint*, arXiv:2512.08296.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. [OpenAssistant Conversations - Democratizing Large Language Model Alignment](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. [Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets](#). *Transactions of the Association for Computational Linguistics*, 10:50–72.

- Ilya Kulikov, Chenxi Whitehouse, Swarnadeep Saha, Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jack Lanchantin, Yoram Bachrach, Jakob Foerster, Xian Li, Han Fang, Sainbayar Sukhbaatar, and Jason Weston. 2026. [Autodata: an automatic data scientist to create high quality data](#).
- Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. [Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. [Tulu 3: Pushing Frontiers in Open Language Model Post-Training](#). In *Second Conference on Language Modeling*.
- Irene Li, Thomas George, Alex Fabbri, Tammy Liao, Benjamin Chen, Rina Kawamura, Richard Zhou, Vanessa Yan, Swapnil Hingmire, and Dragomir Radev. 2023. [A Transfer Learning Pipeline for Educational Resource Discovery with Application in Survey Generation](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 29–43, Toronto, Canada. Association for Computational Linguistics.
- Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. 2024. [Table-GPT: Table Fine-tuned GPT for Diverse Table Tasks](#). *Proc. ACM Manag. Data*, 2(3).
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. [Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. [Chain-of-Dictionary Prompting Elicits Translation in Large Language Models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yansong Tang, and Weizhu Chen. 2024. [Arena Learning: Build Data Flywheel for LLMs Post-training via Simulated Chatbot Arena](#). *Preprint*, arXiv:2407.10627.
- Alisia Lupidi, Carlos Gemmell, Nicola Cancedda, Jane Dwivedi-Yu, Jason Weston, Jakob Foerster, Roberta Raileanu, and Maria Lomeli. 2025. [Source2synth: Synthetic data generation and curation grounded in real data sources](#). *Preprint*, arXiv:2409.08239.
- Xing Han Lù and Siva Reddy. 2026. [Structured distillation of web agent capabilities enables generalization](#). *Preprint*, arXiv:2604.07776.
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. [EuroLLM-9B: Technical Report](#). *Preprint*, arXiv:2506.04079.
- Lester James V. Miranda, Ivan Vulić, and Anna Korhonen. 2026. [Polyglot Teachers: Evaluating Language Models for Multilingual Synthetic Data Generation](#). *Preprint*, arXiv:2604.11290.
- Arindam Mitra, Luciano Del Corro, Guoqing Zheng, Shweti Mahajan, Dany Rouhana, Andres Codash, Yadong Lu, Wei ge Chen, Olga Vrousos, Corby Rosset, Fillipe Silva, Hamed Khanpour, Yash Lara, and Ahmed Awadallah. 2024. [AgentInstruct: Toward Generative Teaching with Agentic Flows](#). *Preprint*, arXiv:2407.03502.
- David Mora, Viraat Aryabumi, Wei-Yin Ko, Sara Hooker, Julia Kreutzer, and Marzieh Fadaee. 2025. [The Art of Asking: Multilingual Prompt Optimization for Synthetic Data](#). *Preprint*, arXiv:2510.19806.

- 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. 2023. [Crosslingual Generalization through Multitask Finetuning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.
- Peter Mwangi. 2020. *Elementary Swahili: Lugha na Utamaduni*. Open Textbook Series. Northwestern University Libraries, Evanston, IL.
- Jaroslav Victor Nigrin. 1918. *Bohemian Grammar*. Slavie Publishing Co., Chicago, IL.
- NVIDIA, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduchintala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek, Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amy Shen, Andrew Tao, and 197 others. 2025. [NVIDIA Nemotron Nano 2: An Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model](#). *Preprint*, arXiv:2508.14444.
- OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. [gpt-oss-120b & gpt-oss-20b Model Card](#). *Preprint*, arXiv:2508.10925.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. [Cross-lingual Name Tagging and Linking for 282 Languages](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.
- Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. [The Berkeley Function Calling Leaderboard \(BFCL\): From Tool Use to Agentic Evaluation of Large Language Models](#). In *Forty-second International Conference on Machine Learning*.
- Juan N. Pava, Thomas S. Mullaney, Caroline Meinhardt, Audrey Gao, and Diyi Yang. 2026. [How Can AI Support Language Digitization and Digital Inclusion?](#) White paper, Stanford University Human-Centered Artificial Intelligence (HAI) and SILICON.
- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrana, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. [Survey of Cultural Awareness in Language Models: Text and Beyond](#). *Computational Linguistics*, 51(3):907–1004.
- José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and Andre Martins. 2025. [M-Prometheus: A Suite of Open Multilingual LLM Judges](#). In *Second Conference on Language Modeling*.
- Dwaipayan Roy, Sumit Bhatia, and Prateek Jain. 2022. [Information asymmetry in wikipedia across different languages: A statistical analysis](#). *J. Assoc. Inf. Sci. Technol.*, 73(3):347–361.
- Alejandro R. Salamanca, Diana Abagyan, Daniel D’souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaei, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan Gomez, Ivan Zhang, and 7 others. 2026. [Tiny Aya: Bridging Scale and Multilingual Depth](#). *Preprint*, arXiv:2603.11510.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. [WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

- Uri Shaham, Jonathan Herzig, Roei Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. [Multilingual Instruction Tuning With Just a Pinch of Multilinguality](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.
- Luisa Shimabucoro, Ahmet Ustun, Marzieh Fadaee, and Sebastian Ruder. 2025. [A Post-trainer’s Guide to Multilingual Training Data: Uncovering Cross-lingual Transfer Dynamics](#). *Preprint*, arXiv:2504.16677.
- Shivalika Singh, Angelika Romanou, Cl  mentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. [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)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Shivalika Singh, Freddie Vargus, Daniel D’souza, B  rje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O’Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemi  nski, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. [Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.
- Earl W. Stevick and Aremu Olaleye. 1963. *Yoruba Basic Course*. U.S. Department of State, Foreign Service Institute, Washington, DC.
- Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. [A Benchmark for Learning to Translate a New Language from One Grammar Book](#). In *The Twelfth International Conference on Learning Representations*.
- Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 49 others. 2026. [Olmo 3](#). *Preprint*, arXiv:2512.13961.
- Katrina Daly Thompson and Antonia Folarin Schleicher. 2001. *Swahili Learners’ Reference Grammar*. African Language Learners’ Reference Grammar Series. National African Language Resource Center (NALRC) Press, University of Wisconsin–Madison, Madison, WI. ISBN: 1-58684-115-7.
- Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanic, Alexan Ayrapetyan, and Igor Gitman. 2025. [OpenMathInstruct-2: Accelerating AI for Math with Massive Open-Source Instruction Data](#). In *The Thirteenth International Conference on Learning Representations*.
- Harsh Trivedi, Tushar Khot, Mareike Hartmann, Ruskin Manku, Vinty Dong, Edward Li, Shashank Gupta, Ashish Sabharwal, and Niranjan Balasubramanian. 2024. [AppWorld: A Controllable World of Apps and People for Benchmarking Interactive Coding Agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076, Bangkok, Thailand. Association for Computational Linguistics.
- David Wadden, Kejian Shi, Jacob Morrison, Alan Li, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2025. [SciRIFF: A Resource to Enhance Language Model Instruction-Following over Scientific Literature](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6072–6109, Suzhou, China. Association for Computational Linguistics.
- Mingyang Wang, Alisa Stoll, Lukas Lange, Heike Adel, Hinrich Schuetze, and Jannik Str  tgen. 2025. [Bring Your Own Knowledge: A Survey of Methods for LLM Knowledge Expansion](#). In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*, pages 150–168, Vienna, Austria. Association for Computational Linguistics.
- Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. [Expanding Pretrained Models to Thousands More Languages via Lexicon-based Adaptation](#). In *Proceedings of the 60th Annual Meeting of the Association*

- for *Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. [Self-Preference Bias in LLM-as-a-Judge](#). *Preprint*, arXiv:2410.21819.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned Language Models are Zero-Shot Learners](#). In *International Conference on Learning Representations*.
- 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.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. 2025.  [\$\tau\$ -bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains](#). In *The Thirteenth International Conference on Learning Representations*.
- Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2024. [LexC-Gen: Generating Data for Extremely Low-Resource Languages with Large Language Models and Bilingual Lexicons](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13990–14009, Miami, Florida, USA. Association for Computational Linguistics.
- Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. [Beyond Counting Datasets: A Survey of Multilingual Dataset Construction and Necessary Resources](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilya Kulikov, Kyunghyun Cho, Yuandong Tian, Jason E Weston, and Xian Li. 2026. [NaturalReasoning: Reasoning in the Wild with 2.8M Challenging Questions](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. [Hire a Linguist!: Learning Endangered Languages in LLMs with In-Context Linguistic Descriptions](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.
- Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. [MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages](#). *Transactions of the Association for Computational Linguistics*, 11:1114–1131.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. [WildChat: 1M ChatGPT Interaction Logs in the Wild](#). In *The Twelfth International Conference on Learning Representations*.
- Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy K Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet S Sekhon, and 7 others. 2026. [Establishing Best Practices in Building Rigorous Agentic Benchmarks](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

## Appendix & Supplementary Material

### A Information About Use of AI Assistants

We used LLMs when preparing the manuscript and writing research code. Specifically, we used Opus 4.6 to improve grammar, usage, and on brainstorming research titles. For research code, we use Opus 4.6 and Codex to optimize the wording of prompts (see Figure 6, Figure 7, and Figure 8) and on debugging errors. All scientific claims and interpretations are solely our own. We reviewed and revised all LLM-assisted text.

### B Computational Budget

All experiments were performed in a cluster containing NVIDIA GH200/H100 chips. In the same chip, data generation via DATASMITH can take 2-3 hours to produce roughly 10k instances. Supervised finetuning takes around an hour for a 12B model using LoRA and 5k instances. Evaluation takes 30 minutes for all three trials.

### C Experimental Setup and Hyperparameters

**Table 6** Hyperparameters for finetuning a 12B student model from Gemma 3 12B Instruct with LoRA.

| Hyperparameter  | Value | Hyperparameter     | Value  |
|-----------------|-------|--------------------|--------|
| Learning rate   | 2e-4  | Batch size         | 32     |
| Epochs          | 2     | Grad. Accum. Steps | 4      |
| Max seq. length | 2,048 | Weight decay       | 0.001  |
| Optimizer       | AdamW | Scheduler          | Linear |

**Finetuning** We use the Unsloth framework (Han et al., 2023) for supervised finetuning. For LoRA, we set  $r = 16$  and  $\alpha = 16$ . We also implement early-stopping based on a held-out validation set (5% of training data). The hyperparameter values are found in Table 6.

**Evaluation** For evaluation, we use lm-eval-harness (Gao et al., 2024; Biderman et al., 2026) for the canonical evaluation setup of Global-MMLU and Global PIQA. We then implement CulturalBench’s easy and hard subsets as custom tasks. When reporting CulturalBench results, we decided on reporting the average of both subsets in order to have a single number. Finally, since CulturalBench’s subsets are divided per language, we map each language to the country where it is widely spoken (Table 7).

**System Prompt**

You are a synthetic-data author. Your job is to produce ONE realistic multi-turn chat dialogue between a curious human and an AI assistant, written entirely in {language}.

You will work in two phases:

1. RESEARCH — actively use the tools provided to you to gather grounding information about the topic before you write anything. Search, browse, look up facts, cross-check sources. You may call tools several times and reason between calls. Tool output may be in English; that is fine — you will translate into {language} when you write the dialogue.
2. WRITE — once you have enough grounded information, emit the full dialogue as structured output (a list of role/content messages).

**LANGUAGE (non-negotiable):**

- The dialogue messages MUST be ENTIRELY in {language}. Every word of every turn is in {language} — no English, not even greetings, transitions, fillers, or loanwords that have a natural {language} equivalent.
- Translate or transliterate proper nouns and technical terms as a native {language} speaker naturally would.
- Your tool calls and reasoning may be in English — only the final emitted dialogue must be in {language}.

**CULTURAL GROUNDING:**

- The human in the dialogue is a native {language} speaker who lives where {language} is spoken. Their references, examples, and interests reflect the culture, geography, sports, food, politics, celebrities, and daily life of {language} speakers — not generic Western/American defaults.
- If the seed topic has a generic or Western framing, localize it naturally when it makes sense (local sports, teams, food, places, examples a real {language} speaker would think of first).

**DIALOGUE SHAPE:**

- Strict alternation: user, assistant, user, assistant, ... starting with the user.
- The user turns are short (1–3 sentences), natural chat-style messages. The human SEEKS information, REACTS to the assistant's answers, or PUSHES BACK on claims. The human NEVER offers help, NEVER provides answers, NEVER summarizes.
- Vary the user's reactions across turns. Pull from this menu (mix freely):
  - \* Follow-up question on a specific detail the assistant mentioned
  - \* Clarification request ("Wait, what do you mean by...")
  - \* Pushback or debate on a questionable claim
  - \* Counter-view from elsewhere ("I read that actually...")
  - \* Personal observation ("In my experience...")
  - \* Request for a concrete example
  - \* Shift to a related but different subtopic
- The assistant turns ANSWER the user's questions with real, substantive information grounded in what the tools returned. Do not deflect, do not just suggest resources, do not respond with only links. Keep replies focused and concise.

**OUTPUT FORMAT (strict):**

- Your final reply MUST be a single JSON object and NOTHING else.
- No prose, no commentary, no markdown code fences, no leading/trailing text.
- The JSON object MUST match exactly this shape:

```
{
  "messages": [
    {"role": "user", "content": "<text in {language}>"},
    {"role": "assistant", "content": "<text in {language}>"},
    ...
  ]
}
```

- Allowed values for "role" are "user" and "assistant" only.
- Messages MUST strictly alternate, starting with "user".
- The "content" of every message is a non-empty string in {language}.

Figure 6 System prompt used in DataSmith for full-trajectory dialogue generation.

### Generation Prompt (DataSmith)

Seed topic for the dialogue (use this ONLY as inspiration for what the human is curious about — do NOT copy or echo it verbatim into any message):

{seed\_prompt}

Generate a complete dialogue with EXACTLY {num\_turns} messages total (alternating user/assistant, starting with the user). The dialogue must be entirely in {language}.

The FIRST user turn must be a fresh, natural chat-style question in {language}, written from the human's point of view — short (1–3 sentences), localized to a native {language} speaker's culture and references where it makes sense. It must NOT be a translation, paraphrase, or recitation of the seed prompt above; treat the seed only as a hint about the topic the human is curious about.

Before writing the dialogue, use your tools to research the topic — search, look things up, gather concrete facts that the assistant can ground its answers in. Reason between tool calls. Once you have enough information, emit the dialogue.

REMINDER: your final reply MUST be a single JSON object with a top-level “messages” array, exactly {num\_turns} entries long, alternating user/assistant. No prose, no markdown fences, no preamble — just the JSON.

**Figure 7** Generation prompt used in DataSmith for full-trajectory dialogue generation. The {seed\_prompt}, {num\_turns}, and {language} placeholders are filled in per sample at generation time.

### Generation Prompt (Teacher-Only)

Seed topic for the dialogue (use this ONLY as inspiration for what the human is curious about — do NOT copy or echo it verbatim into any message):

{seed\_prompt}

Generate a complete dialogue with EXACTLY {num\_turns} messages total (alternating user/assistant, starting with the user). The dialogue must be entirely in {language}.

The FIRST user turn must be a fresh, natural chat-style question in {language}, written from the human's point of view — short (1–3 sentences), localized to a native {language} speaker's culture and references where it makes sense. It must NOT be a translation, paraphrase, or recitation of the seed prompt above; treat the seed only as a hint about the topic the human is curious about.

REMINDER: your final reply MUST be a single JSON object with a top-level “messages” array, exactly {num\_turns} entries long, alternating user/assistant. No prose, no markdown fences, no preamble — just the JSON.

**Figure 8** Generation prompt used in Teacher-only multilingual dialogue generation. The {seed\_prompt}, {num\_turns}, and {language} placeholders are filled in per sample at generation time.

**Table 7 Mapping from evaluation language to CulturalBench country subset.** Since Arabic and Swahili are macrolanguage spoken in different countries with widely different cultures, we decide to not evaluate them on CulturalBench. On the other hand, Tagalog and Cebuano share the Philippines subset.

| Language   | Code | Country Subset |
|------------|------|----------------|
| Arabic     | ara  | —              |
| German     | deu  | Germany        |
| French     | fra  | France         |
| Spanish    | spa  | Spain          |
| Japanese   | jpn  | Japan          |
| Czech      | ces  | Czech Republic |
| Indonesian | ind  | Indonesia      |
| Tagalog    | tgl  | Philippines    |
| Cebuano    | ceb  | Philippines    |
| Swahili    | swa  | —              |
| Yoruba     | yor  | Nigeria        |

**Table 8 Language information.** We use DATASMITH to generate synthetic data for the 11 languages listed above, with resource classifications drawn from Joshi et al. (2020). We indicate whether each language is supported by a translation model (TranslateGemma, Finkelstein et al., 2026). We also mark whether a language appears as an evaluation subset in our downstream benchmarks.

| Language   | ISO 639-3 | Resource Availability | Translation Support | Evaluation Benchmarks               |                                     |                                      |
|------------|-----------|-----------------------|---------------------|-------------------------------------|-------------------------------------|--------------------------------------|
|            |           |                       |                     | Global PIQA<br>(Chang et al., 2025) | Global MMLU<br>(Singh et al., 2025) | CulturalBench<br>(Chiu et al., 2025) |
| Arabic     | ara       | 5 (High)              | ✓                   | ✓                                   | ✓                                   | —                                    |
| German     | deu       | 5 (High)              | ✓                   | ✓                                   | ✓                                   | ✓                                    |
| French     | fra       | 5 (High)              | ✓                   | ✓                                   | ✓                                   | ✓                                    |
| Spanish    | spa       | 5 (High)              | ✓                   | ✓                                   | ✓                                   | ✓                                    |
| Japanese   | jpn       | 5 (High)              | ✓                   | ✓                                   | ✓                                   | ✓                                    |
| Czech      | ces       | 4 (Medium-High)       | ✓                   | ✓                                   | ✓                                   | ✓                                    |
| Indonesian | ind       | 3 (Medium)            | ✓                   | ✓                                   | ✓                                   | ✓                                    |
| Tagalog    | tgl       | 3 (Medium)            | ✓                   | ✓                                   | ✓                                   | ✓                                    |
| Cebuano    | ceb       | 3 (Medium)            | ✓                   | —                                   | —                                   | ✓                                    |
| Swahili    | swa       | 2 (Low-Medium)        | ✓                   | ✓                                   | ✓                                   | —                                    |
| Yoruba     | yor       | 2 (Low-Medium)        | ✓                   | ✓                                   | ✓                                   | ✓                                    |

**Table 9 Grammar books used for DataSmith.** For languages at resource level 4 or below (Joshi et al., 2020), we provide a tool that searches a grammar book. We list the grammar book used for each language along with its license. Most grammar books fall under public domain, while some have permissible licenses.

| Language   | Grammar Book                                                  | License       |
|------------|---------------------------------------------------------------|---------------|
| Cebuano    | Baura et al. (1967)                                           | Public domain |
| Czech      | Nigrin (1918)                                                 | Public domain |
|            | Hilchey et al. (2018)                                         | CC-BY-SA 3.0  |
| Indonesian | Foreign Service Institute (1970)                              | Public domain |
|            | Agustini et al. (2026)                                        | CC-BY-NC 4.0  |
| Swahili    | Thompson and Schleicher (2001)                                | Public domain |
|            | Mwangi (2020)                                                 | CC-BY-SA 4.0  |
| Tagalog    | Bloomfield (1917)                                             | Public domain |
|            | Defense Language Institute, Foreign Language Center (1985b,a) | Public domain |
| Yoruba     | Crowther (1843, 1852)                                         | Public domain |
|            | Stevick and Olaleye (1963)                                    | Public domain |

**Table 10 DataSmith API.** Functions and their keyword arguments exposed to the agent.

| Function                                                                     | Description                                                                                                                                                                                                                                                                                                                                          |
|------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Web</b>                                                                   |                                                                                                                                                                                                                                                                                                                                                      |
| <code>web_search(query, max_results=5)</code>                                | General web search. Returns titles, URLs, and short snippets. Use for finding sources and quick factual lookups where a snippet is enough.                                                                                                                                                                                                           |
| <code>web_news_search(query, max_results=5)</code>                           | Recent news articles. Returns date, source, title, URL, and snippet. Prefer over <code>web_search</code> for time-sensitive information.                                                                                                                                                                                                             |
| <code>web_read_url(url, max_chars=8192)</code>                               | Fetch the full readable body of a page (markdown). Use only when you need the actual article text. Do not call it if a <code>web_search</code> snippet already answers the question.                                                                                                                                                                 |
| <code>wikipedia_search(query, max_results=5)</code>                          | Search Wikipedia for articles matching a query. Returns titles and short snippets. The agent is instructed to phrase the query in English even when the surrounding dialogue is being generated in another language.                                                                                                                                 |
| <code>wikipedia_fetch_page(title, max_chars=8192)</code>                     | Fetch the full plain-text content of a Wikipedia article. Used after <code>wikipedia_search</code> to get the actual article body for grounding; pass the exact title returned by search.                                                                                                                                                            |
| <b>Linguistic Resources</b>                                                  |                                                                                                                                                                                                                                                                                                                                                      |
| <code>linguistic_wiktionary_lookup(word, lang="English")</code>              | Look up a word on Wiktionary for definitions, part of speech, example sentences, and morphology. Searches English Wiktionary ( <a href="https://en.wiktionary.org">en.wiktionary.org</a> ), which has entries for all languages. Set <code>lang</code> to the full language name as it appears on Wiktionary (e.g., "Tagalog", "Swahili", "German"). |
| <code>linguistic_translate_word(word, source_lang="eng", target_lang)</code> | Translate a word using a pre-built bilingual dictionary from PanLex. Uses ISO 639-3 language codes (e.g., "eng", "tgl", "sw"). If the dictionary hasn't been built yet, it will be built automatically in the background on first call.                                                                                                              |
| <code>linguistic_search_grammar(query, lang="tgl", max_results=3)</code>     | Search a pre-indexed grammar reference book for information about the target language. Returns matching passages with page numbers. Use for morphology, syntax, verb conjugation, sentence patterns, etc. Uses BM25 (via <code>bm25s</code> ) over chunks extracted from a PDF.                                                                      |
| <code>linguistic_read_grammar_page(lang="tgl", page)</code>                  | Read a specific page from a grammar index. Use after <code>linguistic_search_grammar</code> to read surrounding context for a relevant passage.                                                                                                                                                                                                      |
| <b>Expert Systems</b>                                                        |                                                                                                                                                                                                                                                                                                                                                      |
| <code>expert_translate_text(text, source_lang="en", target_lang)</code>      | Translate a passage of text using a dedicated translation model. Uses the TranslateGemma prompt format ( <code>&lt;&lt;source&gt;&gt;en&lt;&lt;target&gt;&gt;tgl&lt;&lt;text&gt;&gt;...</code> ). <code>source_lang</code> and <code>target_lang</code> must be ISO 639-1 Alpha-2 language codes (e.g., "tl", "sw").                                 |
| <code>expert_language_id(text, top_k=1)</code>                               | Identify the language of a text passage using GlotLID. Returns an ISO 639-3 code with script (e.g., <code>tgl_Latn</code> , <code>ceb_Latn</code> ) and confidence score. Use to verify generated text is in the target language.                                                                                                                    |
| <code>expert_consult_multilingual_lm(query, lang="English")</code>           | Ask a strong multilingual LM for help. Use for verifying phrasing, checking grammar, suggesting idioms, or getting a second opinion on generated text.                                                                                                                                                                                               |

**Table 11 Full results on  $S_\phi$  = Gemma 3 12B Instruct.** For CulturalBench, we report the average of the easy and hard subsets to obtain a single number per language. Subscript values denote the standard error over three runs.

| <i>High-Resource Languages</i>           |                        |                        |                        |                        |                        |                        |
|------------------------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Method                                   | Arabic                 | German                 | French                 | Spanish                | Japanese               |                        |
| <i>Global MMLU</i> (Singh et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                        | 55.2                   | 67.1                   | 51.7                   | 67.1                   | 62.4                   |                        |
| + DATASMITH (Ours)                       | <b>60.3</b> $\pm 0.42$ | 66.2 $\pm 1.35$        | 65.2 $\pm 1.36$        | <b>70.5</b> $\pm 1.25$ | <b>64.3</b> $\pm 0.37$ |                        |
| + Teacher-only $\Phi_\emptyset$          | 58.6 $\pm 1.72$        | <b>67.8</b> $\pm 2.34$ | <b>68.0</b> $\pm 1.34$ | 68.5 $\pm 0.24$        | 63.6 $\pm 0.30$        |                        |
| + Translate-Train                        | 52.4 $\pm 1.70$        | 58.1 $\pm 1.64$        | 62.4 $\pm 1.55$        | 61.3 $\pm 1.57$        | 54.3 $\pm 1.69$        |                        |
| <i>Global PIQA</i> (Chang et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                        | <b>77.0</b>            | 82.0                   | 82.0                   | 81.0                   | 86.0                   |                        |
| + DATASMITH (Ours)                       | 76.3 $\pm 0.33$        | <b>85.0</b> $\pm 0.58$ | 81.3 $\pm 0.33$        | <b>83.3</b> $\pm 0.67$ | <b>88.3</b> $\pm 0.33$ |                        |
| + Teacher-only $\Phi_\emptyset$          | 76.7 $\pm 0.33$        | 80.7 $\pm 0.33$        | <b>83.7</b> $\pm 0.33$ | 78.0 $\pm 1.15$        | 84.3 $\pm 0.33$        |                        |
| + Translate-Train                        | 54.0 $\pm 1.15$        | 79.3 $\pm 0.33$        | 82.0 $\pm 0.58$        | 75.7 $\pm 0.67$        | 78.0 $\pm 1.15$        |                        |
| <i>CulturalBench</i> (Chiu et al., 2025) |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                        | —                      | 54.7                   | 58.3                   | 68.0                   | 65.1                   |                        |
| + DATASMITH (Ours)                       | —                      | <b>59.4</b> $\pm 0.90$ | <b>64.3</b> $\pm 2.06$ | <b>70.0</b> $\pm 0.72$ | 68.9 $\pm 0.54$        |                        |
| + Teacher-only $\Phi_\emptyset$          | —                      | 56.3 $\pm 0.90$        | 60.7 $\pm 2.06$        | 68.8 $\pm 1.44$        | <b>70.1</b> $\pm 0.31$ |                        |
| + Translate-Train                        | —                      | 49.5 $\pm 0.52$        | 46.4 $\pm 2.06$        | 51.2 $\pm 1.44$        | 48.1 $\pm 1.09$        |                        |
| <i>Low- to Mid-Resource Languages</i>    |                        |                        |                        |                        |                        |                        |
| Method                                   | Czech                  | Indonesian             | Swahili                | Tagalog                | Cebuano                | Yoruba                 |
| <i>Global MMLU</i> (Singh et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                        | 42.2                   | 58.2                   | 60.6                   | 39.7                   | —                      | 26.7                   |
| + DATASMITH (Ours)                       | 61.1 $\pm 0.39$        | <b>67.7</b> $\pm 0.45$ | <b>61.7</b> $\pm 2.48$ | <b>56.0</b> $\pm 0.39$ | —                      | <b>38.0</b> $\pm 1.39$ |
| + Teacher-only $\Phi_\emptyset$          | 56.1 $\pm 0.18$        | 65.8 $\pm 1.33$        | 54.0 $\pm 1.28$        | 52.3 $\pm 0.32$        | —                      | 36.3 $\pm 2.39$        |
| + Translate-Train                        | <b>62.2</b> $\pm 0.28$ | 64.5 $\pm 1.70$        | 59.7 $\pm 1.76$        | 52.8 $\pm 0.28$        | —                      | 29.3 $\pm 1.60$        |
| <i>Global PIQA</i> (Chang et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                        | 79.0                   | 95.0                   | 82.0                   | 81.0                   | —                      | 59.0                   |
| + DATASMITH (Ours)                       | <b>81.0</b> $\pm 0.58$ | <b>97.0</b> $\pm 1.20$ | <b>83.0</b> $\pm 1.15$ | <b>88.0</b> $\pm 1.15$ | —                      | <b>63.0</b> $\pm 0.58$ |
| + Teacher-only $\Phi_\emptyset$          | 78.3 $\pm 0.33$        | 94.0 $\pm 0.58$        | 82.3 $\pm 0.67$        | 75.3 $\pm 1.33$        | —                      | 58.3 $\pm 0.33$        |
| + Translate-Train                        | 66.0 $\pm 0.58$        | 90.3 $\pm 0.33$        | 64.3 $\pm 1.20$        | 78.0 $\pm 1.73$        | —                      | 44.3 $\pm 1.67$        |
| <i>CulturalBench</i> (Chiu et al., 2025) |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                        | 52.0                   | <b>59.6</b>            | —                      | 45.6                   | 45.6                   | 74.5                   |
| + DATASMITH (Ours)                       | 56.0 $\pm 1.15$        | <b>59.6</b> $\pm 1.11$ | —                      | <b>50.0</b> $\pm 1.11$ | <b>52.2</b> $\pm 0.64$ | <b>79.5</b> $\pm 1.31$ |
| + Teacher-only $\Phi_\emptyset$          | <b>58.0</b> $\pm 2.31$ | 57.7 $\pm 1.11$        | —                      | 45.6 $\pm 1.28$        | 48.9 $\pm 1.28$        | 75.8 $\pm 1.52$        |
| + Translate-Train                        | 36.7 $\pm 1.33$        | 46.2 $\pm 0.00$        | —                      | 37.0 $\pm 0.98$        | 31.1 $\pm 0.64$        | 38.6 $\pm 1.31$        |

**Table 12 Full results on  $S_\phi$  = Tiny Aya Global 3B.** For CulturalBench, we report the average of the easy and hard subsets to obtain a single number per language. Subscript values denote the standard error over three runs.

| High-Resource Languages           |                 |                 |                 |                 |                 |                 |
|-----------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Method                            | Arabic          | German          | French          | Spanish         | Japanese        |                 |
| Global MMLU (Singh et al., 2025)  |                 |                 |                 |                 |                 |                 |
| $S_\phi$ (no SFT)                 | 51.5            | 57.2            | 55.7            | 56.7            | 52.7            |                 |
| + DATASMITH (Ours)                | 54.0 $\pm$ 1.02 | 58.5 $\pm$ 0.33 | 60.1 $\pm$ 0.33 | 58.7 $\pm$ 0.67 | 56.2 $\pm$ 0.58 |                 |
| + Teacher-only $\Phi_\emptyset$   | 50.2 $\pm$ 0.33 | 54.5 $\pm$ 1.02 | 56.0 $\pm$ 1.02 | 58.5 $\pm$ 1.02 | 55.5 $\pm$ 1.67 |                 |
| + Translate-Train                 | 49.7 $\pm$ 1.67 | 54.7 $\pm$ 0.88 | 55.5 $\pm$ 0.32 | 58.0 $\pm$ 0.92 | 53.2 $\pm$ 1.45 |                 |
| Global PIQA (Chang et al., 2025)  |                 |                 |                 |                 |                 |                 |
| $S_\phi$ (no SFT)                 | 60.0            | 52.0            | 63.0            | 65.0            | 63.0            |                 |
| + DATASMITH (Ours)                | 73.3 $\pm$ 0.33 | 73.0 $\pm$ 0.58 | 77.3 $\pm$ 0.33 | 76.0 $\pm$ 0.58 | 76.0 $\pm$ 0.58 |                 |
| + Teacher-only $\Phi_\emptyset$   | 64.0 $\pm$ 1.20 | 65.0 $\pm$ 1.45 | 67.0 $\pm$ 1.67 | 74.3 $\pm$ 0.33 | 76.0 $\pm$ 1.45 |                 |
| + Translate-Train                 | 63.0 $\pm$ 1.20 | 56.7 $\pm$ 0.33 | 56.2 $\pm$ 1.53 | 72.0 $\pm$ 0.58 | 72.0 $\pm$ 0.58 |                 |
| CulturalBench (Chiu et al., 2025) |                 |                 |                 |                 |                 |                 |
| $S_\phi$ (no SFT)                 | —               | 40.6            | 39.3            | 45.0            | 57.5            |                 |
| + DATASMITH (Ours)                | —               | 55.7 $\pm$ 0.52 | 56.0 $\pm$ 1.19 | 58.8 $\pm$ 0.72 | 56.6 $\pm$ 0.54 |                 |
| + Teacher-only $\Phi_\emptyset$   | —               | 42.2 $\pm$ 0.90 | 45.2 $\pm$ 1.19 | 40.0 $\pm$ 0.72 | 60.4 $\pm$ 0.54 |                 |
| + Translate-Train                 | —               | 44.8 $\pm$ 0.52 | 32.1 $\pm$ 0.00 | 36.3 $\pm$ 0.00 | 58.5 $\pm$ 0.54 |                 |
| Low- to Mid-Resource Languages    |                 |                 |                 |                 |                 |                 |
| Method                            | Czech           | Indonesian      | Swahili         | Tagalog         | Cebuano         | Yoruba          |
| Global MMLU (Singh et al., 2025)  |                 |                 |                 |                 |                 |                 |
| $S_\phi$ (no SFT)                 | 51.1            | 53.7            | 44.0            | 50.4            | —               | 40.5            |
| + DATASMITH (Ours)                | 58.6 $\pm$ 1.02 | 57.0 $\pm$ 1.02 | 52.5 $\pm$ 0.67 | 51.1 $\pm$ 1.02 | —               | 48.2 $\pm$ 0.02 |
| + Teacher-only $\Phi_\emptyset$   | 54.9 $\pm$ 0.67 | 54.7 $\pm$ 0.33 | 46.0 $\pm$ 0.58 | 47.7 $\pm$ 0.58 | —               | 39.5 $\pm$ 0.02 |
| + Translate-Train                 | 57.4 $\pm$ 0.33 | 54.0 $\pm$ 0.33 | 45.2 $\pm$ 0.67 | 47.5 $\pm$ 0.67 | —               | 42.0 $\pm$ 0.02 |
| Global PIQA (Chang et al., 2025)  |                 |                 |                 |                 |                 |                 |
| $S_\phi$ (no SFT)                 | 67.0            | 82.0            | 70.0            | 61.0            | —               | 59.0            |
| + DATASMITH (Ours)                | 80.3 $\pm$ 0.33 | 86.7 $\pm$ 0.33 | 75.0 $\pm$ 0.58 | 75.7 $\pm$ 0.33 | —               | 81.0 $\pm$ 0.05 |
| + Teacher-only $\Phi_\emptyset$   | 75.0 $\pm$ 1.67 | 85.0 $\pm$ 1.20 | 74.0 $\pm$ 0.58 | 69.0 $\pm$ 0.58 | —               | 70.0 $\pm$ 0.05 |
| + Translate-Train                 | 67.0 $\pm$ 1.67 | 83.3 $\pm$ 0.33 | 74.3 $\pm$ 0.33 | 47.0 $\pm$ 0.58 | —               | 77.1 $\pm$ 0.05 |
| CulturalBench (Chiu et al., 2025) |                 |                 |                 |                 |                 |                 |
| $S_\phi$ (no SFT)                 | 36.0            | 34.6            | —               | 32.2            | 32.2            | 52.3            |
| + DATASMITH (Ours)                | 44.0 $\pm$ 1.15 | 38.5 $\pm$ 1.11 | —               | 34.4 $\pm$ 0.64 | 34.4 $\pm$ 0.08 | 55.0 $\pm$ 1.31 |
| + Teacher-only $\Phi_\emptyset$   | 42.0 $\pm$ 1.15 | 36.5 $\pm$ 0.00 | —               | 27.8 $\pm$ 0.64 | 31.1 $\pm$ 0.64 | 47.7 $\pm$ 1.31 |
| + Translate-Train                 | 32.0 $\pm$ 1.15 | 36.5 $\pm$ 0.00 | —               | 24.1 $\pm$ 0.37 | 28.9 $\pm$ 0.00 | 42.4 $\pm$ 0.76 |

**Table 13 Full results on  $S_\phi = \text{Llama 3.1 8B Instruct}$ .** For CulturalBench, we report the average of the easy and hard subsets to obtain a single number per language. Subscript values denote the standard error over three runs.

| High-Resource Languages           |                        |                        |                        |                        |                        |                        |
|-----------------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| Method                            | Arabic                 | German                 | French                 | Spanish                | Japanese               |                        |
| Global MMLU (Singh et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                 | 44.5                   | 51.5                   | 52.2                   | 60.0                   | 54.2                   |                        |
| + DATASMITH (Ours)                | <b>62.0</b> $\pm 1.03$ | <b>68.2</b> $\pm 0.92$ | <b>67.7</b> $\pm 0.67$ | <b>64.5</b> $\pm 0.67$ | <b>64.7</b> $\pm 1.11$ |                        |
| + Teacher-only $\Phi_\emptyset$   | 51.0 $\pm 1.24$        | 59.0 $\pm 0.22$        | 61.5 $\pm 1.11$        | 61.5 $\pm 0.33$        | 45.7 $\pm 0.67$        |                        |
| + Translate-Train                 | 54.2 $\pm 0.67$        | 54.2 $\pm 1.04$        | 58.0 $\pm 0.48$        | 58.7 $\pm 1.02$        | 50.5 $\pm 0.80$        |                        |
| Global PIQA (Chang et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                 | 55.0                   | 66.0                   | 71.0                   | 72.0                   | <b>78.0</b>            |                        |
| + DATASMITH (Ours)                | <b>64.3</b> $\pm 0.33$ | <b>68.7</b> $\pm 0.33$ | <b>79.0</b> $\pm 0.58$ | <b>82.0</b> $\pm 0.67$ | 70.0 $\pm 0.67$        |                        |
| + Teacher-only $\Phi_\emptyset$   | 54.0 $\pm 1.15$        | 66.0 $\pm 1.45$        | 72.3 $\pm 0.33$        | 73.3 $\pm 0.33$        | 71.3 $\pm 0.33$        |                        |
| + Translate-Train                 | 57.0 $\pm 1.15$        | 47.3 $\pm 1.15$        | 65.7 $\pm 0.33$        | 73.0 $\pm 1.00$        | 59.7 $\pm 0.33$        |                        |
| CulturalBench (Chiu et al., 2025) |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                 | —                      | 48.4                   | 57.1                   | 56.2                   | 65.1                   |                        |
| + DATASMITH (Ours)                | —                      | <b>58.8</b> $\pm 0.52$ | <b>64.1</b> $\pm 1.19$ | <b>65.0</b> $\pm 0.72$ | <b>67.5</b> $\pm 0.31$ |                        |
| + Teacher-only $\Phi_\emptyset$   | —                      | 57.8 $\pm 0.90$        | 58.3 $\pm 2.38$        | 61.3 $\pm 0.72$        | 57.2 $\pm 0.31$        |                        |
| + Translate-Train                 | —                      | 50.0 $\pm 0.90$        | 57.1 $\pm 2.06$        | 55.0 $\pm 0.72$        | 60.4 $\pm 0.54$        |                        |
| Low- to Mid-Resource Languages    |                        |                        |                        |                        |                        |                        |
| Method                            | Czech                  | Indonesian             | Swahili                | Tagalog                | Cebuano                | Yoruba                 |
| Global MMLU (Singh et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                 | 38.3                   | 44.5                   | 37.7                   | 42.4                   | —                      | 25.2                   |
| + DATASMITH (Ours)                | <b>56.3</b> $\pm 1.11$ | <b>58.2</b> $\pm 0.48$ | <b>50.5</b> $\pm 1.11$ | <b>56.5</b> $\pm 0.67$ | —                      | <b>41.0</b> $\pm 0.02$ |
| + Teacher-only $\Phi_\emptyset$   | 50.6 $\pm 0.67$        | 56.0 $\pm 0.52$        | 43.5 $\pm 0.88$        | 46.4 $\pm 1.02$        | —                      | 37.5 $\pm 0.02$        |
| + Translate-Train                 | 47.4 $\pm 0.33$        | 56.2 $\pm 1.02$        | 39.5 $\pm 0.56$        | 46.7 $\pm 0.33$        | —                      | 28.7 $\pm 0.02$        |
| Global PIQA (Chang et al., 2025)  |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                 | 52.0                   | 55.0                   | 56.0                   | 52.0                   | —                      | 38.0                   |
| + DATASMITH (Ours)                | <b>71.0</b> $\pm 0.58$ | <b>86.3</b> $\pm 0.33$ | <b>75.7</b> $\pm 0.33$ | <b>53.0</b> $\pm 0.58$ | —                      | <b>49.0</b> $\pm 0.58$ |
| + Teacher-only $\Phi_\emptyset$   | 62.0 $\pm 1.45$        | 84.0 $\pm 0.04$        | 73.0 $\pm 1.15$        | 49.3 $\pm 0.33$        | —                      | 39.0 $\pm 0.58$        |
| + Translate-Train                 | 64.0 $\pm 0.67$        | 83.3 $\pm 0.33$        | 59.0 $\pm 1.67$        | <b>53.7</b> $\pm 0.33$ | —                      | 45.0 $\pm 1.67$        |
| CulturalBench (Chiu et al., 2025) |                        |                        |                        |                        |                        |                        |
| $S_\phi$ (no SFT)                 | 40.0                   | 53.8                   | —                      | 47.8                   | 47.8                   | 50.0                   |
| + DATASMITH (Ours)                | <b>46.0</b> $\pm 1.15$ | 51.9 $\pm 1.11$        | —                      | <b>52.2</b> $\pm 0.64$ | <b>56.7</b> $\pm 0.64$ | <b>59.1</b> $\pm 1.31$ |
| + Teacher-only $\Phi_\emptyset$   | 44.0 $\pm 1.15$        | 50.0 $\pm 1.11$        | —                      | 42.2 $\pm 0.64$        | 46.7 $\pm 0.64$        | 52.3 $\pm 1.31$        |
| + Translate-Train                 | 36.0 $\pm 1.15$        | <b>55.8</b> $\pm 1.11$ | —                      | 43.3 $\pm 0.64$        | 42.2 $\pm 0.00$        | 56.8 $\pm 1.31$        |

## D LLM-as-a-judge Experiment Details

**Annotating reference trajectories** To calibrate the LLM judge, we manually annotate a set of existing agent trajectories using the interface shown in Figure 9. Judging research trajectories does not require fluency in the target language. The authors therefore annotated 100 trajectories to build a pool of gold reference judgments. The resulting inter-annotator agreement with the LM judge is reported in Table 5 in the main text.

**Prompting the Judge LM** We use GPT-OSS-120B (Apache 2.0, OpenAI et al., 2025) as our LLM judge. We choose a different model family to prevent self-preference bias (Wataoka et al., 2025; Chen et al., 2025) which can confound the results. We follow the recommended sampling parameters with temperature of 1.0, top-k of 0, and top-p of 1.0. The actual prompt is shown in Figure 10.

**Example of a Low Research-Efficiency Trajectory** Figure 11 shows an excerpt of a Yoruba research trajectory. The teacher issues a long sequence of single-sentence translation calls, one for each sentence in the seed prompt, before writing the dialogue, illustrating the lower research efficiency we observe for low-resource languages (§6.2).

**Score the quality of DataSmith Trajectories**

**Instructions:** Read the Research Phase (the agent's thinking and tool calls) and the Generation Phase (the resulting dialogue), then score every dimension below and click Save & next.

**Tagalog**

**Seed:** Can you give me questions to address that are related to bullying?

**Research Phase**

tool\_call: web\_search

tool\_result: web\_search

tool\_call: web\_read\_url

tool\_result: web\_read\_url

URL: <https://www.abonibai.com/the-situpule-against-bullying-in-chilipine-schools/>

[Facebook 1] [X 2]

**Generation Phase**

Kapatid ko pataging umiyak pag-uwi galing eskwelahan. Sabi niya may grupo raw na lagi siyang tinutukso at inasser. Paano ko kaya siya matutulungan?

Nakakaawa naman ang kapatid mo. Ang sitwasyon na iyan ay halimbawa ng pang-aapi o bullying. Ayon sa Batas Republika 10627 o Anti-Bullying Act of 2013, ipinagbabawal ang anumang paulit-ulit na pag-atake—pisikal, verbal, o online—na nakakasira sa kanyang kalooban. Mahalagang ipaalam mo sa magulang ninyo ang nangyayari para magreport sa paaralan. Lahat ng paaralan sa Pilipinas ay may tungkuling magkaroon ng patakaran laban sa pang-aapi, kasama ang paraan ng pag-uulat at counseling.

0 / 10 annotated · Item 2 of 10 this session

**Evidence Groundedness**

Are the claims and content of the synthesized dialogue are supported by information the agent actually retrieved through tools?

Score  
1 = very poor ... 5 = excellent

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Justification (optional)

Notes (optional)

**Information Density**

Does a large proportion of the agent's retrieved information relevant and useful for producing the sample?

Score  
1 = very poor ... 5 = excellent

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Justification (optional)

**Research Efficiency**

Is the agent efficient in using tool calls during the whole trajectory? Does it constantly repeat tool-use calls in an irrelevant manner?

Score  
1 = very poor ... 5 = excellent

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Justification (optional)

Use via API · Built with Gradle · Settings

**Figure 9 Reference annotation interface.** Screenshot of the annotation UI used to collect human-annotated scores for reference research trajectories, which serve as few-shot exemplars when prompting the LM judge.

**Judge System Prompt**

You are an expert evaluator of research trajectories produced by a tool-using agent. The agent synthesizes multilingual instruction-tuning data: each trajectory is a sequence of reasoning, external tool calls, and tool outputs, ending in a synthesized instruction-response dialogue. You grade the PROCESS of the trajectory and how well it grounds that dialogue. In this domain, “retrieved information” means external tool outputs.

Score each of the following dimensions from 1 to 5 (5 = best), evaluating each dimension INDEPENDENTLY and based ONLY on the trajectory and synthesized dialogue provided. Anchors: 1 = very poor, 2 = poor, 3 = adequate, 4 = good, 5 = excellent.

1. **evidence\_groundedness**: Are the claims and content of the synthesized dialogue supported by information the agent actually retrieved through tools?
2. **information\_density**: Does a large proportion of the agent’s retrieved information relevant and useful for producing the sample?
3. **research\_efficiency**: Is the agent efficient in using tool calls during the whole trajectory? Does it constantly repeat tool-use calls in an irrelevant manner?

Return a STRICT JSON object and nothing else. For each dimension provide an integer “score” (1–5) and a one-sentence “justification”. Use exactly these keys: `evidence_groundedness`, `information_density`, `research_efficiency`.

**Judge Prompt**

Task (synthesis spec): <SEED PROMPT>  
Target language: <LANGUAGE>

Trajectory:  
[1] THINKING: <thinking>  
[2] TOOL CALL `web_search({"query": "<query>"})`  
[3] TOOL RESULT `web_search`: <retrieved text>

Synthesized dialogue:  
USER: <user turn>  
ASSISTANT: <assistant turn>

Score the trajectory and synthesized dialogue above on each rubric dimension. Return only the JSON object.

**Figure 10 Prompts used for the LLM-as-a-judge evaluation of research trajectories.** The system prompt (top) defines the three scoring rubrics and the strict JSON output format, while the judge prompt (bottom) supplies the trajectory and synthesized dialogue to be scored.

Thinking

The user wants a Yoruba dialogue about logical-reasoning questions, localized rather than a translation of the seed. I will use `expert_translate_text` to render the example sentences into Yoruba before writing the dialogue. [... trimmed ...]

Tool Call

```
expert_translate_text({"text": "A woman with pink-striped hair is about to bite a piece of food...", "source_lang": "en", "target_lang": "yo"})
```

Tool Result

Obìnrin kan tí o ní irun tí o ní àwọ̀n àlǎ pínkì [...]

Tool Call

```
expert_translate_text({"text": "Spiky hair watches.", "source_lang": "en", "target_lang": "yo"})
```

Tool Result

Àwọ̀n ọ̀ṣùnkàn tí o ní irun tí o dá.

Tool Call

```
expert_translate_text({"text": "Three men are on stage with their guitars...", "source_lang": "en", "target_lang": "yo"})
```

Tool Result

Ọ̀kùnrin m'ẹ̀ta wà lórí ìgbé, wọ̀n sì w'ọ̀ p'ẹ̀lú àwọ̀n gítàrà [...]

Tool Call

```
expert_translate_text({"text": "A band is performing for a small crowd.", "source_lang": "en", "target_lang": "yo"})
```

Tool Result

Egb'ẹ̀ orin kan ní ẹ̀ ẹ̀se ìdánìl'ẹ̀j'ọ̀ fún àwọ̀n ènìyàn kéré.

Tool Call

```
expert_translate_text({"text": "A man celebrates a child's first birthday.", "source_lang": "en", "target_lang": "yo"})
```

Tool Result

Ọ̀kùnrin kan ẹ̀se ayẹyẹ ọ̀j'ọ̀ ìbí àkọkọ ọmọ r'ẹ̀.

Tool Call

```
expert_translate_text({"text": "His child was born yesterday.", "source_lang": "en", "target_lang": "yo"})
```

Tool Result

Ọmọ rẹ̀ bí ní ànǎ.

(... 10 more single-sentence `expert_translate_text` calls, one per seed sentence ...)

Thinking

The seed topics involve determining if a premise leads to a hypothesis. I will have a Yoruba speaker ask about such logical conclusions and explain the reasoning. [... trimmed ...]

**Figure 11 Excerpt of a Yoruba research trajectory.** Before writing the dialogue, the teacher makes 16 separate single-sentence `expert_translate_text` calls (using a TranslateGemma expert), one for each sentence in the seed prompt and most of them unused in the final output. This pattern of tool use exemplifies the lower research efficiency we observe for low-resource languages (§6.2).