THINK MULTILINGUAL, NOT HARDER: A FRAMEWORK FOR ANALYZING AND TEACHING CODE-SWITCHED REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent developments in reasoning capabilities have enabled large language models to solve increasingly complex mathematical, symbolic, and logical tasks. Interestingly, while reasoning models are often trained to generate monolingual text, these models have also been observed to code-switch (i.e., mix languages). Prior works have either viewed code-switching as an undesirable error, attempted to control code-switching through modifications to input prompts or the output decoding process, or focus on narrow subsets of languages, domains, tasks, and models. We address these gaps by introducing the first linguistically and behaviorally motivated fine-tuning framework for identifying beneficial code-switched reasoning behaviors in large language models and teaching these models to code-switch more effectively for reasoning. We create the Code-Switched Reasoning (CoRe) corpus, consisting of (1) 7k reasoning traces from 15 models, 18 languages, 10 scripts, and diverse reasoning domains, providing insights into potentially helpful code-switching behaviors, and (2) 40 carefully curated datasets for training and evaluating six interventions for improving code-switching in reasoning across three models and seven languages, totaling 120 fine-tuning conditions. Across 80k+ reasoning traces from both language/culture-agnostic and -specific evaluations, English-dominated reasoning, semantically accurate code-switching, and a more even mix of languages are positively associated with correct answers, whereas dense switching is associated with incorrect answers. Moreover, we are able to elicit positive behaviors through fine-tuning tasks that do not directly demonstrate code-switching. Our work suggests that small but well-curated datasets can change how reasoning models code-switch, allowing us to reap the benefits of reasoning in the many languages that lack large-scale reasoning data.

1 INTRODUCTION

When human multilinguals mix multiple languages during communication, they are said to *code-switch* (Myers-Scotton, 2017). Recently, works including Guo et al. (2025) have observed the parallel phenomenon of code-switching in *reasoning*—the long chain-of-thought texts produced by large language models (LLMs) when solving complex problems. While mechanisms of human code-switching have been studied extensively (e.g., see Myers-Scotton (2017) and Bullock & Toribio (2009)), in comparison, the emergent behavior of code-switching in LLM reasoning remains relatively poorly understood. Some works (e.g., Guo et al., 2025; Marchisio et al., 2024)) treat code-switching as an error, while others attempt to use code-switching strategically for reasoning tasks in particular (Li et al., 2025a;b). In this work, we are particularly interested in code-switched LLM reasoning as a route to improve reasoning capabilities for lower-resource languages. This issue is of critical importance given that currently available reasoning models are mostly limited to high-resource languages like English (Ghosh et al., 2025).

To understand how state-of-the-art reasoning models code-switch and design interventions that improve code-switching quality, we seek answers to the following questions:

RQ 1 How do LLMs code-switch during reasoning?

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

RQ 2 What types of code-switching are associated with better performance on tasks requiring reasoning?

RQ 3 What types of fine-tuning interventions can teach reasoning models beneficial forms of code-switching for reasoning?

To address these questions, we first create and analyze a corpus of reasoning examples from diverse models, languages, tasks, and domains (section 3). Based on the findings from this initial corpus, we design supervised fine-tuning approaches (including training data and evaluation on parallel and non-parallel language/culture-specific and -agnostic benchmarks) to develop beneficial code-switched reasoning behaviors (section 4).

Our contributions are as follows. First, we introduce the **Code-Switched Reasoning (CoRe)** corpus of 7k reasoning traces from diverse models, languages, tasks, and domains. With this initial corpus, we develop and apply a taxonomy of code-switching behaviors to identify characteristics that benefit reasoning performance. Second, we introduce a linguistically and behaviorally motivated fine-tuning framework for teaching reasoning models to code-switch for better reasoning performance. This framework extends **CoRe** with 1M tokens of training data per task per language, across 6 tasks and 7 languages. Third, we systematically analyze code-switching behaviors produced by our framework and their relationship with reasoning performance. We find that fine-tuning for translation tasks most consistently shifts code-switching toward behaviors associated with correct answers. Crucially, this benefit appears for low resource languages. Overall, we hope that this work will serve as the foundation for further future work on code-switching as a means to improving multilingual LLM reasoning in low-resource settings. We will release all reasoning traces, SFT data, trained adapters, and evaluation scripts needed to reproduce the main figures under an MIT License upon acceptance.

2 Related Work

We incorporate prior insights from research on human code-switching to contribute to a growing body of work on controlling the languages that models use during reasoning (e.g., Tam et al., 2025; Gao et al., 2025). Broadly defined, code-switching occurs when a multilingual person mixes multiple languages during communication (Myers-Scotton, 2017). Human code-switching is driven by a complex array of social and cognitive factors, e.g., ethnolinguistic identity and language proficiency (Myers-Scotton, 2017; Gumperz, 1977; Bullock & Toribio, 2009). While language models—and particularly reasoning models—have likewise been observed to code-switch, the mechanisms behind this behavior remain poorly understood. Some works treat code-switching by language models as an undesirable behavior (Marchisio et al., 2024; Guo et al., 2025). However, our work more closely aligns with Li et al., who attempt to harness code-switching as a useful reasoning behavior. The main difference between this work and Li et al. (2025a) is that while they target the decoding stage, we modify model behavior through supervised fine-tuning.

Prior work on code-switched reasoning lacks our coverage of diverse models, languages, domains, tasks, and reasoning behaviors. Yong et al. (2025) study the English-centric s1 model family using commonsense, factual, cultural, and causal reasoning benchmarks, with prompts in four languages other than English. Li et al. (2025b;a) study Chinese/English math reasoning from the Chinese-English QwQ32B-Preview model. Wang et al. (2025b;a) study the Chinese-English DeepSeek-R1 model family and multilingual QwQ-32B, Qwen3, and Gemini 2.0 Flash Thinking models on commonsense, factual, and logical reasoning in 15 languages. In contrast, we cover 17 models from 10 families with diverse multilingual capabilities, 21 languages, and reasoning domains/tasks beyond those in prior work, e.g., moral reasoning (taking into consideration the experiments and analyses from both section 3 and section 4). When characterizing code-switching behaviors, Yong et al. identify the "quote-and-think" pattern and apply the pre-existing linguistic concepts of a matrix language and intra-/inter-sentential switching. Wang et al. quantify relative proportions of reasoning languages. Li et al. identify four main patterns. In contrast, we introduce a structured taxonomy of code-switching patterns across three dimensions and quantify these patterns through 9 code-switching metrics validated by Linguistics and NLP literatures.

Our work combines top-down, theory-driven and bottom-up, data-driven approaches for classifying reasoning behavior. Gandhi et al. (2025) identify reasoning behaviors that both drive perfor-

mance and parallel human behaviors, while Lee et al. (2025) characterize reasoning through bottom-up LLM-assisted brainstorming and clustering. We build on Lee et al. but focus on code-switching, add manual curation to ground the taxonomy in linguistic theories of human code-switching, and include smaller multilingual models and domains beyond math, science, and general knowledge.

In contrast with prior work that controls reasoning language at inference time, we modify code-switching behavior through supervised fine-tuning (SFT). Yong et al. (2025) and Zhao et al. (2026) steer reasoning language with prompt instructions and by prefilling target-language tokens; Li et al. (2025a) train a probe to guide language choice during decoding; and Wang et al. (2025a) mask the logits of non-target scripts. We treat these prior inference-time approaches as potentially complementary to our SFT approach, rather than as baselines. Our SFT approach is partially inspired by Muennighoff et al. (2025a;b), who show that reasoning capabilities can be developed with SFT on a small, carefully curated corpus.

3 Understanding Code-Switched Reasoning Behaviors

To answer **RQ 1** on how LLMs code-switch, we create a new corpus, **CoRe**. We begin with an initial subset of 7k reasoning traces from 15 models, described as follows.

3.1 Approach

Both top-down and bottom-up approaches to characterizing reasoning behaviors (cf. section 2) face limitations. Theory-grounded, top-down approaches which look for pre-defined, human-aligned behaviors in model reasoning (e.g., Gandhi et al.) may miss novel reasoning behaviors that do not align with humans. On the other hand, data-driven, bottom-up approaches which derive categories of behaviors from observations of model reasoning (e.g., (Lee et al., 2025)) may fail to surface misalignment with categories of human behaviors. In particular, these data-driven, bottom-up approaches may fail to observe behaviors that only appear in humans, and not in models. To address these limitations, we fuse top-down theory-driven and bottom-up data-driven approaches.

Data and models. To create our **CoRe** corpus, we generate examples of reasoning from diverse models, languages, tasks, and domains for generalizability. We generate reasoning examples from 15 models of diverse sizes, reasoning capabilities, and multilingual capabilities (see appendices A.1 and A.3 for full list). We select prompts for generating reasoning examples from seven datasets (see appendix A.3 for full list). Prompts cover 18 languages (Arabic, Bengali, Burmese, English, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Mandarin Chinese, Portuguese, Russian, Spanish, Swahili, Thai, and Yoruba), 10 scripts, and 8 language families (Hammarström et al., 2025; Unicode, 2025). We include approximately 50 instances from each model/language/dataset combination, yielding 7k instances total. (See appendix A.4 for the prompting configurations used to generate the reasoning examples.)

Code-switched reasoning behavior taxonomy. Next, we develop a taxonomy of code-switched reasoning behaviors by fusing top-down theory-driven and bottom-up data-driven approaches. We rely on the reasoning examples described above as input context to Gemini 2.5 Flash for brainstorming criteria that differentiate code-switching strategies (Google, 2025a; Gemini Team et al., 2025). At this stage, we introduce helpful inductive biases to guide the taxonomy development by providing Gemini 2.5 Flash with a minimal and broad definition of code-switching ("use of multiple scripts or languages"). We apply topic modeling to Gemini's brainstormed criteria, using the BERTopic pipeline (Grootendorst, 2022). Finally, we manually consolidate redundant and overly narrow topics (e.g., "Interpretation of Yoruba Terms") and group topics into dimensions, categories, and subcategories. This manual curation is guided by our knowledge of the relevant literature on code-switching and reasoning in both humans and LLMs. See appendices A.5 and A.6 for details.

Our taxonomy of code-switched reasoning behaviors features three dimensions, grounded both in theories from prior literature and our dataset:

1. **Function** describes the purpose served by a code-switch during reasoning. Common function include *translating* from one language into another and *quoting* material from the initial user prompt in its original language, potentially while reasoning primarily in a different language. Humans also code-switch to quote material in its original language, pro-

vide context in another language, and repeat phrases in another language for emphasis and clarity (Begum et al., 2016; Belani & Flanigan, 2023). Model behavior parallels compensatory code-switching in humans, which occurs when a multilingual switches to one language to compensate for a lack of proficiency in another language (Schächinger Tenés et al., 2023). By code-switching to reason about content from lower-resource languages in higher-resource languages, models potentially mitigate the language resource gap.

2. **Form** describes the structure of code-switching within the reasoning. For example, conditioned on a prompt in one language, a model may switch into another language for a single word, phrase, sentence, or even the entire reasoning trace. Code-switching may occur with varying frequency/density within the reasoning. Additionally, a particular language may take on the role of the *matrix* or main language of the reasoning (Bullock et al., 2018).

3. **Coherence** describes how *fluent* (natural and understandable) and *accurate* (using semantically appropriate terms) the code-switching is. Kuwanto et al. (2024) also find *fluency* and *accuracy* to be relevant concepts for their human evaluation of automatically generated code-switched text.

Human validation of LLM annotations. Our validation approach follows Lee et al. (2025), who achieve more interpretable, comprehensive analyses than prior approaches. We randomly sample 100 instances from the code-switching criteria brainstormed by Gemini 2.5 Flash, then manually annotate them as detailed in appendix A.7. We find that Gemini 2.5 Flash is an effective annotator, generating plausible descriptions of code-switching 85% of the time.

Code-switching metrics. From the datasets listed in appendix A.3 (but excluding any instances already used to develop the taxonomy), we include about 50 instances from each model/dataset/language combination (excluding English, since we find that code-switching is rare when prompted in this language). We use Qwen3.5-27B for instance-level LID to identify which reasoning instances feature code-switching, before computing additional code-switching metrics for those instances (see A.13 for details). Below, we describe the code-switching metrics we use to quantify the aspects of code-switching from our taxonomy. We select metrics that have been previously validated and are widely used by the code-switching research community (Srivastava & Singh, 2021). Many of these metrics rely on word-level language identification (LID; see appendix A.8 for details). We validate both word- and instance-level LID on 140 held-out reasoning traces, obtaining 0.933 word-level and 0.942 instance-level accuracy.

We quantify the **Form** dimension of our taxonomy with five widely used span-based metrics. The *Code-mixing Index (CMI)* measures the degree of code-switching using the proportion of tokens from the dominant language in the text, with a higher CMI indicating a higher degree of code-switching (Das & Gambäck, 2014; Srivastava & Singh, 2021). Given n tokens (including u language-independent tokens) in N languages, if w_i is the number of tokens from the ith language, then CMI $= \frac{\sum_{i=1}^{N}(w_i) - max\{w_i\}}{n-u}$.

The *Multilingual Index (M-Index)* also measures the degree of code-switching, by comparing the relative proportions of tokens from each language present in the text (Barnett et al., 2000; Srivastava & Singh, 2021). A higher M-Index indicates a higher degree of code-switching. Given a text of n tokens in N languages where w_i is the number of tokens from the ith language, M-Index $= \frac{1-\sum_{i=1}^{N}(\frac{w_i}{n})^2}{(N-1)(\sum_{i=1}^{N}(\frac{w_i}{n})^2}$.

Memory measures the correlation in lengths of consecutive spans of tokens in different languages. Higher memory indicates that when a span of tokens in one language is followed by a span of tokens in another language, the two consecutive token spans tend to be of more similar length (Goh & Barabási, 2008; Guzmán et al., 2017; Srivastava & Singh, 2021). Given a text of n_r language spans (where a language span consists of consecutive tokens in the same language), where τ_i denotes the ith language span, σ_1, μ_1 denote the standard deviation and mean of the lengths of all language spans except the last, and σ_2, μ_2 denote the standard deviation and mean of the lengths of all language spans except the first, Memory $= \frac{1}{n_r - 1}\sum_{1}^{n_r-1}\frac{(\tau_i - \mu_1)(\tau_{i+1} - \mu_2)}{\sigma_1 \sigma_2}$.

Burstiness measures how predictable code-switches are, with higher burstiness indicating less predictable code-switching (Srivastava & Singh, 2021; Goh & Barabási, 2008). Given a text consisting

of several spans of tokens in different languages, where μ, σ are the mean and standard deviation of the length of the token spans, Burstiness $= \frac{\sigma-\mu}{\sigma+\mu}$.

The *Integration-Index (I-Index)* measures the probability (over the entire text) that the language of the next token will differ from the language of the previous token. A higher I-Index indicates higher probability of switching languages (Guzmán et al., 2017; Srivastava & Singh, 2021). Given a text of n tokens and S where $S(i, i+1) = 1$ if the ith token is in a different language from the $i+1^{th}$ token, I-Index $= \frac{\sum_{i=1}^{n-1} S(i,i+1)}{n-1}$.

The *matrix language* is the language that provides the grammatical structure of a code-switched sentence, and into which words from a second "embedded" language are inserted (Myers-Scotton, 2001). Adopting the heuristic used by Yong et al. (2025), we identify the language which contributes the most tokens to a reasoning trace as the matrix language. We categorize the matrix language as (1) same as the prompt, (2) English, or (3) other.

To measure *code-switching fluency and accuracy*, we adapt the approach from Kuwanto et al. (2024) for evaluating naturalistic code-switched text generation and use Qwen/Qwen3.5-27B as our rater (see Appendices A.9 and A.10 for details), selected because it is among the strongest open models supporting the low-resource languages in our study.

Statistical modeling. To estimate how code-switching behaviors relate to reasoning performance, we fit a logistic mixed model with lme4 (Bates et al., 2015). We model the correctness y_i of response i to question q in prompt language ℓ as $\text{logit}\, P(y_i{=}1) = \beta_0 + \sum_k \beta_k b_{ik} + \boldsymbol{\gamma}^\top \mathbf{x}_i + u_\ell + v_q$, where u_ℓ and v_q are random intercepts for language and question. The behaviors b_{ik} are indicators for switching language at the prompt–reasoning boundary, switching at the reasoning–answer boundary, code-switching within the reasoning (c_i), and the matrix language. Each remaining metric enters as $c_i z_{ik}$, where z_{ik} is the metric z-scored among code-switched responses. The controls $\mathbf{x}_i$ are dataset and reasoning model. Together, the controls and random intercepts keep estimated effects of code-switching from being trivially explained by differences in overall model performance, language resource level, or question difficulty. We drop the CMI because it is collinear with the M-Index.

3.2 Results

LLM code-switching in reasoning varies widely depending on the prompt language (RQ 1). In Figure 2 (left), we characterize these differences along the dimensions of our taxonomy.

Function. Code-switching often translates prompt content into a higher-resource language or quotes prompt material in its original language (see Figure 1 for an example in Burmese [my]). This parallels compensatory code-switching in humans, in which multilinguals switch to a more proficient language to compensate for weaker proficiency in another language (Schächinger Tenés et al., 2023).

Form. Chinese prompts elicit Chinese as the reasoning matrix language, reflecting the strong Chinese-English capabilities of models such as the DeepSeek-R1 family (Guo et al., 2025). Notably, Arabic and Hindi prompts also frequently elicit Chinese as the matrix language. English, the main language supported by most reasoning models, is a common reasoning matrix language across many prompt languages.

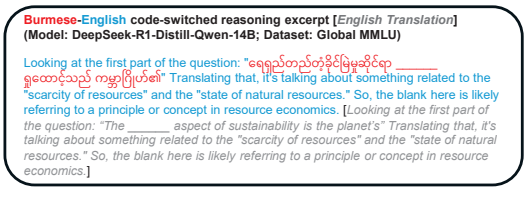

Figure 1: Example of code-switching for the function of translation.

Coherence. Reasoning in response to prompts from high-resource languages such as Chinese (zh) features more semantically accurate code-switching than reasoning in response to prompts from low-resource languages such as Yoruba (yo), likely reflecting the relative proportions of these languages in models' training data. However, accuracy and fluency can diverge. For example, reasoning in response to Burmese prompts is rated as highly fluent but less accurate. This suggests that models can produce smoothly integrated code-switched text that is nonetheless semantically incorrect.

In response to **RQ 2**, we find that reasoning with higher code-switching accuracy is associated with better performance (Figure 2, right). Surprisingly, increased fluency of code-switching, which

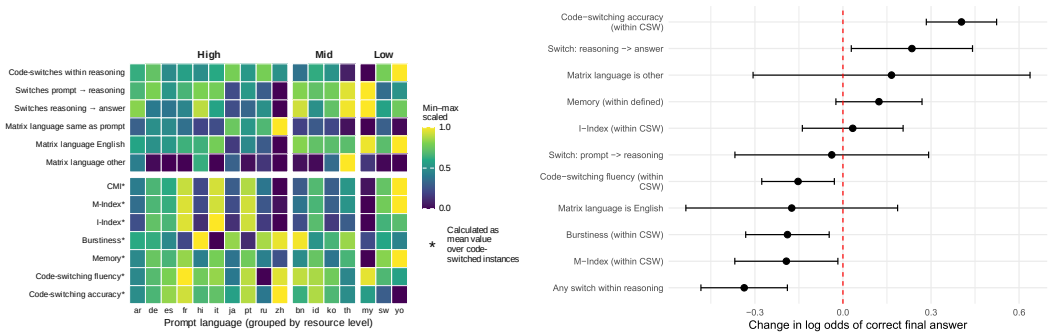

Figure 2: **Left:** Code-switching behaviors in reasoning differ widely by prompt language, along dimensions of **Form**, **Function**, and **Coherence**. **Right:** Code-switching accuracy is positively associated with answer correctness. Location of switching also matters; switches in language when transitioning from the reasoning to the answer phase of the model response is positively associated with final answer correctness, whereas the binary presence/absence of any code-switches within the reasoning phase is negatively associated. (Data is the 7k subset of CoRe described in section 3.1. "Within CSW" refers to metrics that are only defined when there is code-switching within the reasoning part of a model's response, as opposed to switching into/out of a language at the beginning and end of reasoning. Points are fixed-effect estimates from a mixed model with random effects for model and language; bars are 95% CIs.)

implies smoother syntactic integration and flow between languages, is not helpful for reasoning performance. Taken together, these results suggest that the *semantic correctness of* code-switched reasoning is more predictive of performance than its *syntactic integration*.

4 Producing Code-Switching Behaviors Using Linguistically and Behaviorally Motivated Data Curation

Based on our observations of code-switched reasoning behaviors in section 3, we seek to answer **RQ 3** by testing the effect of six fine-tuning interventions on code-switching. By eliciting behavioral changes through these controlled interventions, we are able to move beyond the observational approach in section 3 and instead systematically induce specific code-switched reasoning behaviors predictive of better reasoning performance.

4.1 Approach

Datasets. To extend our **CoRe** corpus for reasoning SFT and evaluation, we use the Global MMLU, Global PIQA, and No Language Left Behind datasets (Singh et al., 2025; Chang et al., 2026).

Global MMLU extends the Massive Multitask Language Understanding (MMLU) dataset to 42 languages, using human translations and human-validated machine translations (Singh et al., 2025; Hendrycks et al., 2021). We select Global MMLU as a source of high-quality reasoning problems from geographically and typologically diverse low-, mid-, and high-resource languages of different families. In particular, we train and evaluate on the Amharic, Hindi, Igbo, Indonesian, Malay, Swahili, and Yoruba subsets of Global MMLU. These seven languages represent three scripts, three resource levels, four language families, and over 1 billion speakers worldwide. By selecting this diverse subset of languages, we can observe whether our fine-tuning interventions generalize across languages with different characteristics (Singh et al., 2025; Eberhard et al., 2026).

Recently, Sprague et al. (2025) found that reasoning primarily benefits model performance in math, science, symbolic reasoning, legal argument reasoning, and moral reasoning. Therefore, we filter Global MMLU to only consider questions related to these domains and tasks. We borrow Sprague et al.'s heuristic and include all questions that feature "=" in the question text. Since Global MMLU also categorizes questions by subject, we also curate a list of reasoning-related subjects and include all questions from these subjects. (See appendix A.11 for the full list.) From the resulting Global MMLU instances, we set aside validation and test sets of 500 examples each.

Global PIQA evaluates commonsense reasoning in $100+$ languages using questions written by native speakers (Chang et al., 2026). In addition to evaluating on a curated subset of examples from Global MMLU (as described above), we evaluate our fine-tuned models on the non-parallel split of Global PIQA, which evaluates culturally specific commonsense reasoning. The Global PIQA evaluation allows us to test how well our fine-tuning interventions generalize beyond the translated benchmarks (like Global MMLU), to non–English-centric settings. We evaluate in the same seven languages as our Global MMLU subset, using the full 100-instance Global PIQA non-parallel test set for each language.

No Language Left Behind (NLLB) created open-source data and models for translation between over 200 languages (Costa-jussà et al., 2022). We select NLLB for fine-tuning on our machine translation task (see below for more details), because of its extensive multilingual coverage and demonstrated usefulness in developing state-of-the-art machine translation models.

Models. We apply our fine-tuning interventions to three reasoning models, representing different parameter sizes, model families, and multilingual capabilities. *Qwen3-8B* is an eight-billion parameter model supporting 119 language varieties. *Phi-4-Reasoning* is a 14-billion parameter model that only fully supports English. *DeepSeek-R1-Distill-Llama-8B* is an eight-billion parameter model with strong capabilities in English and Chinese (Guo et al., 2025).

To generate fine-tuning data for those experimental conditions featuring reasoning demonstrations, we prompt Qwen3-Next-80B-A3B-Thinking using the default recommended sampling parameters (Yang et al., 2025). We select Qwen3-Next-80B-A3B-Thinking as the teacher model for its strong multilingual reasoning capabilities. Additionally, to test the generalizability of our approach to other teacher models, we also generate demonstrations using gemma-4-26B-A4B-it and fine-tune the Qwen 3-8B student model on these demonstrations in Swahili. We leave a full-scale teacher generalization test to future work.

Some of our interventions use machine translation (MT) as part of the data synthesis pipeline, for which we use the SeamlessM4T (Barrault et al., 2023), NLLB-200 (Costa-jussà et al., 2022), and 70B Apertus models (Apertus et al., 2025). We provide details of our MT pipeline in appendix A.12.

Supervised fine-tuning tasks. The general format for all supervised fine-tuning tasks is (p, r, a): conditioned on prompt p, the reasoning model is fine-tuned to generate reasoning r leading to final answer a. To ensure a fair comparison across tasks and languages, we set a token budget of 1 million tokens for fine-tuning each model in each language for a given task. Because the same content tokenizes into different numbers of tokens in different languages (Ahia et al., 2023), a fixed token budget yields different numbers of training examples per language. We therefore control for the number of training examples in our downstream analyses. Due to constraints on the amount of high-quality data we are able to generate for Yoruba, we do not release datasets for the native-language reasoning or synthetically code-switched reasoning tasks in this language (see below for explanation of these tasks). Full details of our fine-tuning configuration are in appendix A.14.

We summarize the six tasks below (l and e denote the target language and English, respectively). Appendix A.18 gives full data synthesis details.

Native-language reasoning (p_l, r_l, a_l) is a monolingual baseline. We generate reasoning demonstrations from the teacher model, then filter for appropriate length and correctness before translating the reasoning into l.

Machine translation into English $(p_l, \emptyset, a_e)$ trains sentence-level translation on NLLB with empty reasoning. It targets the translation function of code-switching (section 3) without directly demonstrating reasoning.

Reasoning prompt translation into English $(p_l, \emptyset, a_e)$ trains translation of Global MMLU's parallel prompts into English. Like the machine translation task above, it targets the translation function of code-switching observed in section 3.

English-language reasoning (p_l, r_e, a_l) switches into English for the entire reasoning trace. This task is motivated by analysis of code-switching **Form** in section 3, which suggests that the location and granularity of code-switches (e.g., within v. across boundaries between the prompt, reasoning, and answer) are related to final answer accuracy.

Strategically code-switched reasoning (p_l, r_{csw}, a_l) uses the teacher's own code-switched reasoning on p_l, filtered for examples of reasoning containing multiple languages. This task tests the view that code-switching serves a strategic **function** and that a strong multilingual reasoner like our Qwen3-Next-80B-A3B-Thinking teacher model can generate examples of "good" code-switches for reasoning (Li et al., 2025a).

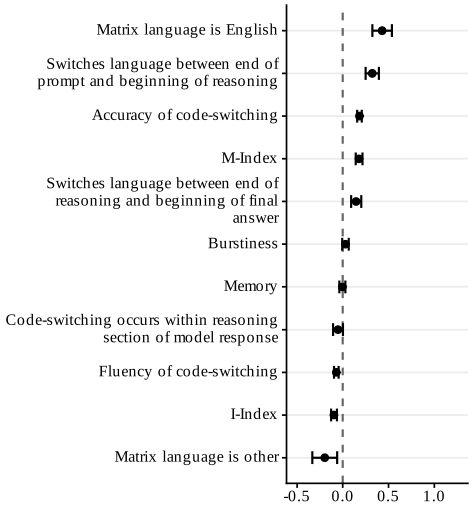

Figure 3: Relationship between code-switching behaviors and change in log odds of correct final answer for models treated by our SFT interventions and evaluated on Global MMLU and Global PIQA.

Synthetically code-switched reasoning (p_l, r_{csw}, a_l) splices together randomly selected steps of r_e and r_l. We include this task to test the effect of the random, and hence less **coherent**, code-switches in contrast with the strategic code-switching task above.

Statistical modeling of relationship between code-switching and performance. On test-set outputs from Global MMLU and Global PIQA, we fit two sets of mixed models with lme4 (Bates et al., 2015), once on the pooled data and once per dataset. First, we model the correctness y_i of response i to question q in language ℓ as logit $P(y_i{=}1) = \beta_0 + \sum_k \beta_k b_{ik} + \boldsymbol{\gamma}^\top \mathbf{x}_i + u_\ell + v_q$, with random intercepts u_ℓ and v_q. The behaviors b_{ik} are indicators for switching language at the prompt–reasoning boundary, switching at the reasoning–answer boundary, and code-switching within the reasoning (c_i), plus the matrix language (English or other, relative to the prompt language). Each remaining metric (section 3.1) enters as $c_i z_{ik}$, where z_{ik} is the metric z-scored among code-switched responses. The controls $\mathbf{x}_i$ are dataset (Global PIQA or Global MMLU), reasoning model, number of SFT examples,[1] and MT quality. We drop the CMI from the regression because it is collinear with the M-Index (as indicated by the adjusted GVIF).

Statistical modeling of the effect of SFT tasks on code-switching behaviors. Second, we estimate how each SFT task shifts each code-switching behavior. For behavior k in response i, we fit $g(\mathbb{E}[b_{ik}]) = \alpha_{k0} + \sum_t \alpha_{kt} \mathbb{1}[\text{task}_i = t] + \boldsymbol{\delta}_k^\top \mathbf{x}'_i + u_\ell + v_q$, where g is the logit link for binary behaviors (e.g., whether English is the matrix language) and the identity link for continuous metrics (e.g., CMI). The base (non-fine-tuned) model is the reference level, so α_{kt} is the change in behavior k associated with SFT task t. The covariates $\mathbf{x}'_i$ match those in the first model, except that they omit the number of SFT examples, which is collinear with task. The terms u_ℓ and v_q are random intercepts for language and question.

4.2 Results

While the results of section 3 (**RQ 2**) inform the design of our SFT interventions, we do not expect identical outcomes in the evaluation of our fine-tuned models, due to differences in the models, languages, and datasets covered, as well as the different random effects structures for modeling the effect of interventions on the code-switching behaviors and the association of code-switching behaviors with accuracy. Therefore, we revisit **RQ 2** before addressing **RQ 3**.

Code-switching behaviors associated with correct answers (RQ 2). Figure 3 shows the correctness model from section 4.1 fit to 80,532 test-set responses from the three base models and their 120 fine-tuned variants. Several associations hold on both benchmarks. Responses whose matrix language is English are more likely to be correct than responses dominated by the prompt language

[1]With a fixed token budget, different languages have different numbers of training examples as discussed above (Ahia et al., 2023).

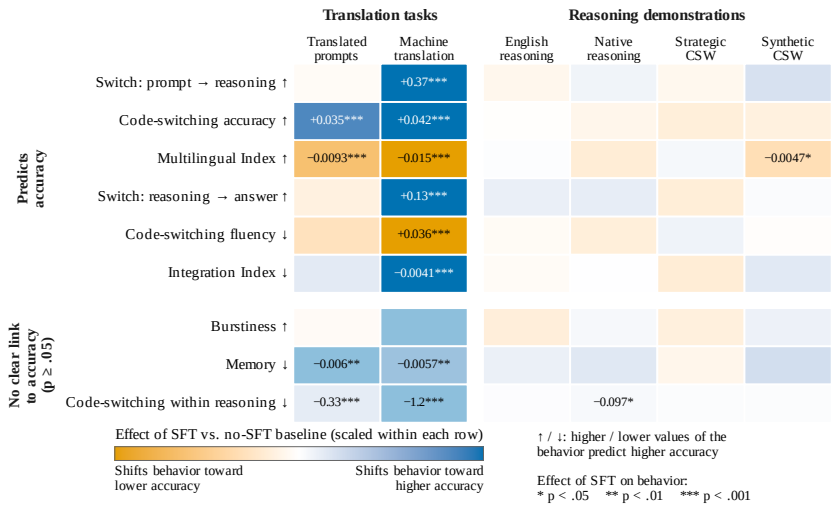

Figure 4: Among our SFT interventions to improve code-switched reasoning, translation-based tasks have the most significant positive effects on code-switching behaviors associated with final answer accuracy. (The numbers in the cells marked as statistically significant represent the effect sizes of the SFT interventions prior to row-wise scaling.)

(pooled β = 0.43, 95% CI [0.32, 0.54]). Among code-switched responses, more semantically accurate code-switching (β = 0.18) and a more balanced mix of languages (M-Index, β = 0.18) are associated with correct answers, whereas a high density of code-switching points (I-Index, β = -0.10) and higher fluency ($\beta = -0.07$) are associated with incorrect answers. Switching language at the reasoning–answer boundary, i.e., reasoning at least partially in one language and answering in another, is also positively associated with correctness on both benchmarks (pooled β = 0.15). See appendix A.19 for more results.

Effects of fine-tuning on code-switching behaviors (RQ 3). Figure 4 shows the causal effect of each intervention on each behavior relative to the base models. Fine-tuning for machine translation, which contains no reasoning demonstrations, has the broadest effects. On the pooled data, it increases code-switching accuracy, lowers the I-Index, and increases the odds of switching language at the prompt–reasoning (+0.37 log odds) and reasoning–answer (+0.13) boundaries. All of these are shifts toward behaviors associated with correct answers ($p < .001$). Fine-tuning to translate reasoning prompts into English likewise increases code-switching accuracy ($p < .001$). In contrast, fine-tuning on reasoning demonstrations (English, native-language, strategically code-switched, or synthetically code-switched reasoning) produces few or no significant changes in any behavior. Translation, rather than reasoning demonstration, thus appears to be the most effective lever for changing how models code-switch, consistent with translation being a central function of code-switching in reasoning (section 3.2).

Generalization to Gemma teacher model. Student models still benefit from demonstrations generated by Gemma with our framework, showing that it has the potential to apply to multiple different teacher models. See appendix A.21 for plots.

5 Conclusion

We address the gap in our understanding of how LLMs code-switch during reasoning, the types of code-switching that are helpful for reasoning, and the types of fine-tuning interventions that can teach good code-switched reasoning, by introducing a framework including data, code, and models for identifying and fine-tuning for these behaviors. We demonstrate that characteristics of code-switching in reasoning such as the matrix language, semantic accuracy, and degree of code-switching all matter for performance. We show that our framework can significantly increase helpful aspects of code-switching for reasoning performance. Our work establishes code-switching as a means for improving reasoning capabilities in languages with limited data resources.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, help developing theoretical models and conceptual frameworks, designing and providing feedback on research methodology and experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, or proposing or refining hypotheses, and the aforementioned tasks are not applicable to this work. Additionally, we used generative AI tools for creating and modifying scientific figures and images, suggesting experimental parameters, creating and editing software code, creating artifacts, brainstorming, sourcing/searching for information, editing the research paper to improve readability, identifying relevant literature, and proposing the title and keywords for the research paper. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness, synthesized data was manually reviewed for quality, and cited literature was human-reviewed before incorporation into the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproduction of our work, we will release our CoRe dataset along with all trained adapteres and evaluation scripts needed to reproduce the results and figures in this work, under an MIT License upon acceptance. Additionally, we have included the LLM prompts we used for data synthesis and analysis in the appendix, and explained details of model hyperparameters/configurations and regression setups in both the main text of the paper and the appendix. We also provide more detailed explanations of our overall data synthesis process in the appendix.

REFERENCES

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

Anthropic. Claude 3.7 sonnet and claude code, 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin

Jaggi, and Imanol Schlag. Apertus: Democratizing open and compliant llms for global language environments, 2025. URL https://arxiv.org/abs/2509.14233.

Ruthanna Barnett, Eva Codó, Eva Eppler, Montse Forcadell, Penelope Gardner-Chloros, Roeland van Hout, Melissa Moyer, Maria Carme Torras, Maria Teresa Turell, Mark Sebba, Marianne Starren, and Sietse Wensing. Automatic analysis. *International Journal of Bilingualism*, 4(2):198–208, 2000. doi: 10.1177/13670069000040020107. URL https://doi.org/10.1177/13670069000040020107.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. Seamless: Multilingual expressive and streaming speech translation, 2023. URL https://arxiv.org/abs/2312.05187.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. Functions of code-switching in tweets: An annotation framework and some initial experiments. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1644–1650, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1260/.

Ritu Belani and Jeffrey Flanigan. Automatic identification of code-switching functions in speech transcripts. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7438–7448, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.469. URL https://aclanthology.org/2023.findings-acl.469/.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Prasoon Varshney, Makesh Narsimhan, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Shaona Ghosh, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Chris

Alexiuk, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, and Eric Chung. Llama-nemotron: Efficient reasoning models, 2025. URL https://arxiv.org/abs/2505.00949.

Barbara Bullock, Wally Guzmán, Jacqueline Serigos, Vivek Sharath, and Almeida Jacqueline Toribio. Predicting the presence of a matrix language in code-switching. In Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg (eds.), *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pp. 68–75, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-3208. URL https://aclanthology.org/W18-3208/.

Barbara E. Bullock and Almeida Jacqueline Toribio. *The Cambridge Handbook of Linguistic Code-switching*. Cambridge University Press, 2009.

Tyler A. Chang, Catherine Arnett, Abdelrahman Sadallah, Abdelrahman Eldesokey, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarimayum Meerajita Sharma, Aditi Gupta, Adril Putra Merin, Adwoa Bremang, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akriti Kuri, Akshay Ramesh, Aleksei Dorkin, Alfred Malengo Kondoro, Alham Fikri Aji, Ali Eren Çetintaş, Allan Hanbury, Alou Dembele, Alp Niksarli, Álvaro Arroyo, Amin Bajand, Amol Khanna, Ana Chkhaidze, Ana Carolina Condez, Anamaria-Roberta Hartl, Andiswa Mkhonto, Andrew Hoblitzell, Andrew Tran, Angelos Poulis, Anirban Majumder, Anjali Chaudhary, Anna Vacalopoulou, Annette Kuuipolani Kanahele Wong, Annika Simonsen, Anton Kovalev, Anupam Nayak, Ashvanth S, Ayodeji Lana, Ayu Purwarianti, Bashar Alhafni, Benedict Busole, Bernard Ghanem, Bharti Nathani, Biljana Stojanovska Đurić, Blessing Ogundipe, Bolaotan Agbonile, Bragi Bergsson, Bruce Torres Fischer, Burak Tutar, Burcu Çınar, Cade Kane, Can Udomcharoenchaikit, Chadi Helwe, Chaithra Reddy Nerella, Chen Cecilia Liu, Chiamaka Nwokolo, Christopher Homan, Clément Sampebgo, Cristina España-Bonet, Cynthia Amol, Daeyoep Lee, Dan Saattrup Smart, Dana Arad, Daniil Dzenhaliou, Dasol Choi, David Liu, David Semedo, David Anugraha, Deborah Popoola, Deividas Mataciunas, Delphine Nyaboke, Dennis Owusu, Dhyuthy Krishna Kumar, Diogo Tavares, Diogo Glória-Silva, Divyanshu Goyal, DongGeon Lee, E. Kelly Buchanan, Ebele Nwamaka Anajemba, Egonu Ngozi Grace, Elena Mickel, Elias Herranen, Eliza Acharya, Eman Nisar, Emile Anand, Emmanuel Habumuremyi, Emuobonuvie Maria Ajiboye, Eryawan Presma Yulianrifat, Esther Adenuga, Ewa Rudnicka, Faith Itiola, Faran Taimoor Butt, Fareeha Fayyaz Sheikh, Fathima Thekkekara, Fatima Haouari, Faustin Nsengiyumva, Fenal Ashokbhai Ilasariya, Filbert Aurelian Tjiaranata, Firas Laakom, Francesca Grasso, Francesco Periti, Francesco Orabona, Gbenga Kayode Solomon, Genta Indra Winata, Gia Nghia Ngo, Gloria Udhedhe-oze, Gonçalo Vinagre, Gopi Naga Sai Ram Challagolla, Gorka Urbizu-Garmendia, Gouthami Vadithya, Guijin Son, Gulnaz Abdykadyrova, Gyan Swaroop Mohapatra, Hafeez Ullah, Hafsteinn Einarsson, Hai Hu, Hamidreza Saffari, Hamza Zaidi, Haopeng Zhang, Harethah Abu Shairah, Harry Vuong, Hele-Andra Kuulmets, Hitesh Laxmichand Patel, Houda Bouamor, Hwanjo Yu, Iben Nyholm Debess, İbrahim Ethem Deveci, Ikhlasul Akmal Hanif, Ikhyun Cho, Inês Vieira, Inês Calvo, Isaac Manzi, Ismael Illa Salifou, Ismail Daud, Ismail Yusuf, Itay Itzhak, Ivan Zhelyazkov, Ivan Belashkin, Ivan Spada, Jacob Brinton, Jafar Isbarov, Jaka Čibej, Jan Kocoń, Jan Cuhel, Jauza Krito, Jebish Purbey, Jennifer Za, Jennifer Mickel, Jenny Kunz, Jessica Ratovondranto, Jeyarajalingam Varsha, Jihae Jeong, Jimena Tena Dávalos, Jinu Lee, João Magalhães, John Seon Keun Yi, Jongin Kim, Joseph Chataignon, Joseph Marvin Imperial, Jubeerathan Thevakumar, Judith Land, Julia Alekseenko, Junchen Jiang, Jungwhan Kim, Kairit Sirts, Kamesh R, Kamesh V, Kanda Tshinu, Kätriin Kukk, Kaustubh Ponkshe, Kavsar Huseynova, Ke He, Kenneth Enevoldsen, Kent Joshua Alvarez, Kerem Zaman, Khalil Mrini, Kian Kyars, Komal Gour, Krishnakumar Lainitha, Krister Kruusmaa, Kunal Mukherjee, Kusum Chouhan, Laura Castro, Laura M. Porrino-Moscoso, Lenny Sivi Za Nzambi, Leshem Choshen, Levent Sencan, Lilja Øvrelid, Lisa Alazraki, Loretta Oma Jones, Lovina Ehimen-Ugbede, Luheerathan Thevakumar, Luxshan Thavarasa, Mahnoor Malik, Mamadou K. Keita, Mansi Jangid, Marco De Santis, Marcos Garcia, Marek Šuppa, Mariam D'Ciofalo, Marii Ojastu, Marium Attaullah, Maryam Sikander, Mausami Narayan, Maximos Skandalis, Mehak Mehak, Mehmet İlteriş Bozkurt, Melaku Bayu, Menan Velayuthan, Mhasilenuo Vizo, Michael Leventhal, Michał Marcińczuk, Mina Almasi, Mirna Potočnjak, Mithil Bangera, Mohammadamin Shafiei, Mohiba Ansari, Mridul Sharma, Mrityunjaya Indoria, Mughees Ur Rehman, Muhammad Ravi Shulthan Habibi, Murat

Kolić, Murat Barkın Kınay, Nada Galant, Naina Singh Rathore, Naphat Permpredanun, Narada Maugin, Nathalie Norman, Nicholas Kluge Corrêa, Nikola Ljubešić, Nirmal Thomas, Nisansa de Silva, Nisheeth Joshi, Nitish Ponkshe, Nizar Habash, Nneoma Udeze, Noel Thomas, Noémi Ligeti-Nagy, Nouhoum Coulibaly, Odunayo Ogundepo, Odunayo Kareemat Buliaminu, Oghojafor Godswill Fejiro, Okechukwu God'spraise, Olanrewaju Samuel, Olaoye Deborah Oluwaseun, Olasoji Akindejoye, Olga Snissarenko, Onyinye Anulika Chiemezie, Orkun Kınay, Osman Tursun, Oyelade Oluwafemi Joshua, Oyesanmi Fiyinfoluwa, Pablo Rodríguez, Pablo Gamallo, Palak Arora, Pedro Valente, Peter Rupnik, Philip Oghenesuowho Ekiugbo, Prakhar Agarwal, Pramit Sahoo, Prokopis Prokopidis, Pua Niau-Puhipau, Quadri Yahya, Rachele Mignone, Raghav Singhal, Rahul Raja, Ram Mohan Rao Kadiyala, Raphael Merx, Rasmus Larsen, Ratnavel Rajalakshmi, Rishav Ghosh, Romina Oji, Ron Kekeha Solis, Rui Guerra, Rushikesh Zawar, Sa'ad Nasir Bashir, Saeed Alzaabi, Sahil Sandeep, Sai Pavan Batchu, Sai Sandeep Kantareddy, Saleha Muzammil, Salsabila Zahirah Pranida, Sam Buchanan, Samuel Rutunda, Sander Land, Sarah Sulollari, Sardar Ali, Saroj Sapkota, Sarveswaran Kengatharaiyer, Saulius Tautvaisas, Sayambhu Sen, Sayantani Banerjee, Sebastien Diarra, Segun Afolayan, Senthilnathan M, Sewoong Lee, Shaan Shah, Shankar Venkitachalam, Sharifa Djurabaeva, Sharon Ibejih, Shivanya Shomir Dutta, Siddhant Gupta, Silvia Paniagua Suárez, Sina Ahmadi, Sivasuthan Sukumar, Siyuan Song, Snegha A, Sokratis Sofianopoulos, Sona Elza Simon, Sonja Benčina, Sophie Gvasalia, Sphurti More, Spyros Dragazis, Stefan Milosavljević, Stephan P. Kaufhold, Suba S, Sultan Alrashed, Surangika Ranathunga, Taiga Someya, Taja Kuzman Pungeršek, Tal Haklay, Tasi'u Jibril, Tatsuya Aoyama, Tea Abashidze, Terenz Jomar Dela Cruz, Terra Blevins, Themistoklis Nikas, Theresa Idoko, Thu Mai Do, Tilek Chubakov, Tina Munda, Tobiloba Owoeye, Tommaso Gargiani, Uma Rathore, Uni Johannesen, Uwuma Ugwu, Vallerie Alexandra Putra, Vanya Bannihatti Kumar, Varvara Arzt, Vasily Konovalov, Vasudevan Nedumpozhimana, Viktoria Ondrejova, Viktoryia Horbik, Vishnu Vardhan Reddy Kummitha, Vuk Dinić, Walelign Sewunetie, Winston Wu, Xiaojing Zhao, Yacouba Diarra, Yaniv Nikankin, Yash Mathur, Yash Bagla, Yeshil Bangera, Yixi Chen, Yiyuan Li, Yolanda Xavier, Yonatan Belinkov, Zaid Alyafeai, Zhargal Batozargalova, Zhengyang Shan, Zhi Rui Tam, Zilu Tang, Zuzana Nadova, Baber Abbasi, Stella Biderman, David Stap, Duygu Ataman, Fabian Schmidt, Hila Gonen, Jiayi Wang, and David Ifeoluwa Adelani. Global piqa: Evaluating commonsense reasoning across 100+ languages and cultures, 2026. URL https://arxiv.org/abs/2510.24081.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. Iterative translation refinement with large language models. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz (eds.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pp. 181–190, Sheffield, UK, June 2024. European Association for Machine Translation (EAMT). URL https://aclanthology.org/2024.eamt-1.17/.

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. Seal: Steerable reasoning calibration of large language models for free, 2025. URL https://arxiv.org/abs/2504.07986.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation, 2022. URL https://arxiv.org/abs/2207.04672.

Amitava Das and Björn Gambäck. Identifying languages at the word level in code-mixed Indian social media text. In Dipti Misra Sharma, Rajeev Sangal, and Jyoti D. Pawar (eds.), *Proceedings of the 11th International Conference on Natural Language Processing*, pp. 378–387, Goa, India, December 2014. NLP Association of India. URL https://aclanthology.org/W14-5152/.

David M. Eberhard, Gary F. Simons, and Alison J. Robinson. Ethnologue: Languages of the world. https://www.ethnologue.com/, 2026. Accessed: 2026-03-18.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars, 2025. URL https://arxiv.org/abs/2503.01307.

Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. Could thinking multilingually empower llm reasoning?, 2025. URL https://arxiv.org/abs/2504.11833.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul

Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick

Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic,

Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. A survey of multilingual reasoning in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8920–8936, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.474. URL https://aclanthology.org/2025.findings-emnlp.474/.

K.-I. Goh and A.-L. Barabási. Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002, jan 2008. doi: 10.1209/0295-5075/81/48002. URL https://doi.org/10.1209/0295-5075/81/48002.

Google. Gemini models, 2025a. URL https://ai.google.dev/gemini-api/docs/models/.

Google. google/gemma-3-4b-it, 2025b. URL https://huggingface.co/google/gemma-3-4b-it.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

John J. Gumperz. The sociolinguistic significance of conversational code-switching. *RELC Journal*, 8(2):1–34, 1977. doi: 10.1177/003368827700800201. URL https://doi.org/10.1177/003368827700800201.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. Metrics for Modeling Code-Switching Across Corpora. In *Interspeech 2017*, pp. 67–71, 2017. doi: 10.21437/Interspeech.2017-1429.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. Glottolog 5.2, 2025. URL https://doi.org/10.5281/zenodo.15525265.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. GlotScript: A resource and tool for low resource writing system identification. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7774–7784, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.687/.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. Big-bench extra hard, 2025. URL https://arxiv.org/abs/2502.19187.

Shivani Kumar and David Jurgens. Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with UniMoral. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5890–5912,

Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.294. URL https://aclanthology.org/2025.acl-long.294/.

Garry Kuwanto, Chaitanya Agarwal, Genta Indra Winata, and Derry Tanti Wijaya. Linguistics theory meets llm: Code-switched text generation via equivalence constrained large language models, 2024. URL https://arxiv.org/abs/2410.22660.

Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, Moontae Lee, and Minjoon Seo. The cot encyclopedia: Analyzing, predicting, and controlling how a reasoning model will think, 2025. URL https://arxiv.org/abs/2505.10185.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. The impact of language mixing on bilingual LLM reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32531–32548, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1654. URL https://aclanthology.org/2025.emnlp-main.1654/.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. The impact of language mixing on bilingual llm reasoning, 2025b. URL https://arxiv.org/abs/2507.15849.

Kelly Marchisio, Wei-Yin Ko, Alexandre Berard, Théo Dehaze, and Sebastian Ruder. Understanding and mitigating language confusion in LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6653–6677, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.380. URL https://aclanthology.org/2024.emnlp-main.380/.

Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, Teven Le Scao, Yihan Wang, Adam Yang, Alexander H. Liu, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Andy Ehrenberg, Anmol Agarwal, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, Clémence Lanfranchi, Darius Dabert, Devon Mizelle, Diego de las Casas, Elliot Chane-Sane, Emilien Fugier, Emma Bou Hanna, Gauthier Delerce, Gauthier Guinet, Georgii Novikov, Guillaume Martin, Himanshu Jaju, Jan Ludziejewski, Jean-Hadrien Chabran, Jean-Malo Delignon, Joachim Studnia, Jonas Amar, Josselin Somerville Roberts, Julien Denize, Karan Saxena, Kush Jain, Lingxiao Zhao, Louis Martin, Luyu Gao, Lélio Renard Lavaud, Marie Pellat, Mathilde Guillaumin, Mathis Felardos, Maximilian Augustin, Mickaël Seznec, Nikhil Raghuraman, Olivier Duchenne, Patricia Wang, Patrick von Platen, Patryk Saffer, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Romain Sauvestre, Rémi Delacourt, Sanchit Gandhi, Sandeep Subramanian, Shashwat Dalal, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Thibault Schueller, Thibaut Lavril, Thomas Robert, Thomas Wang, Timothée Lacroix, Valeriia Nemychnikova, Victor Paltz, Virgile Richard, Wen-Ding Li, William Marshall, Xuanyu Zhang, and Yunhao Tang. Magistral, 2025. URL https://arxiv.org/abs/2506.10910.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset, 2025. URL https://arxiv.org/abs/2504.16891.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025a. Association for Computational

Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Workshop on Reasoning and Planning for Large Language Models*, 2025b. URL https://openreview.net/forum?id=LdH0vrgAHm.

Carol Myers-Scotton. *Codeswitching Worldwide II*, chapter The matrix language frame model: Developments and responses, pp. 23–58. Mouton de Gruyter, Berlin, 2001.

Carol Myers-Scotton. *Code-Switching*, chapter 13, pp. 217–237. John Wiley & Sons, Ltd, 2017. ISBN 9781405166256. doi: https://doi.org/10.1002/9781405166256.ch13. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405166256.ch13.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, and Pattarawat Chormai. PyThaiNLP: Thai natural language processing in Python, June 2024. URL https://github.com/PyThaiNLP/pythainlp/.

L. T. Schächinger Tenés, J. C. Weiner-Bühler, L. Volpin, A. Grob, K. Skoruppa, and R. K. Segerer. Language proficiency predictors of code-switching behavior in dual-language-learning children. *Bilingualism: Language and Cognition*, 26(5):942–958, 2023. doi: 10.1017/S1366728923000081.

Shivalika Singh, Angelika Romanou, Clémentine 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, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=w6nlcS8Kkn.

Vivek Srivastava and Mayank Singh. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan (eds.), *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pp. 6–14, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.calcs-1.2. URL https://aclanthology.org/2021.calcs-1.2/.

Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. Language matters: How do multilingual input and reasoning paths affect large reasoning models?, 2025. URL https://arxiv.org/abs/2505.17407.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Weixin Xu, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles,

Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, Zonghan Yang, and Zongyu Lin. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Unicode. Languages and scripts, 2025. URL https://www.unicode.org/cldr/charts/47/supplemental/languages_and_scripts.html.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. Language mixing in reasoning language models: Patterns, impact, and internal causes. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2637–2665, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.132. URL https://aclanthology.org/2025.emnlp-main.132/.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schütze. Language mixing in reasoning language models: Patterns, impact, and internal causes, 2025b. URL https://arxiv.org/abs/2505.14815.

Rongsheng Wang. The dataset distilled from kimi-k1.5. https://huggingface.co/datasets/wangrongsheng/Kimi-K1.5-Distill-data, 2025.

Haoran Xu, Baolin Peng, Hany Awadalla, Dongdong Chen, Yen-Chun Chen, Mei Gao, Young Jin Kim, Yunsheng Li, Liliang Ren, Yelong Shen, Shuohang Wang, Weijian Xu, Jianfeng Gao, and Weizhu Chen. Phi-4-mini-reasoning: Exploring the limits of small reasoning language models in math, 2025. URL https://arxiv.org/abs/2504.21233.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, Felix Juefei-Xu, Foutse Khomh, Osamu Yoshie, Qingyu Chen, Douglas Teodoro, Nan Liu, Randy Goebel, Lei Ma, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation, 2025. URL https://arxiv.org/abs/2503.10497.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Zheng-Xin Yong, M. Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji. Crosslingual reasoning through test-time scaling, 2025. URL https://arxiv.org/abs/2505.05408.

Raoyuan Zhao, Yihong Liu, Hinrich Schuetze, and Michael A. Hedderich. A comprehensive evaluation of multilingual chain-of-thought reasoning: Performance, consistency, and faithfulness across languages. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5223–5247, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. doi: 10.18653/v1/2026.findings-eacl.276. URL https://aclanthology.org/2026.findings-eacl.276/.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

A Appendix

A.1 Reasoning models used to develop code-switched reasoning behavior taxonomy

Model	Citation
DeepSeek-R1 and its distilled variants	Guo et al. (2025)
gemini-2.0-flash-thinking-exp-1219	Google (2025a); Gemini Team et al. (2025)
Claude Sonnet 3.7	Anthropic (2025)
Llama-3.1-Nemotron-Nano-8B-v1	Bercovich et al. (2025)
gemma-3-4b-it	Google (2025b)
Phi-4-mini-reasoning	Xu et al. (2025)
QwQ-32B	Yang et al. (2024); Team (2025)
Magistral Small 1.0	Mistral-AI et al. (2025)
Kimi k1.5	Team et al. (2025)

A.2 Languages supported by models used to develop code-switched reasoning behavior taxonomy

Model	Supported languages	Unsupported languages
DeepSeek-R1-Distill-Qwen-1.5B	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
DeepSeek-R1-Distill-Qwen-7B	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
DeepSeek-R1-Distill-Llama-8B	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
DeepSeek-R1-Distill-Qwen-14B	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
DeepSeek-R1-Distill-Qwen-32B	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
DeepSeek-R1-Distill-Llama-70B	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
DeepSeek-R1	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru

Model	Supported languages	Unsupported languages
Llama-3.1-Nemotron-Nano-8B-v1	en, de, fr, it, pt, es, th, hi	zh, ar, ja, ko, bn, sw, id, yo, my, ru
phi-4-mini-reasoning	en	zh, es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
gemma-3-4b-it	140+ languages[2]	Unknown
magistral-small-2506	en, fr, de, el, hi, id, it, ja, ko, ms, ne, pl, pt, ro, ru, sr, es, tr, uk, vi, ar, bn, zh, fa	sw, yo, my, th
gemini-2.0-flash-thinking-exp-1219	ar, bn, bg, zh, hr, cs, da, nl, en, et, fi, fr, de, el, iw, hi, hu, id, it, ja, ko, lv, lt, no, pl, pt, ro, ru, sr, sk, sl, es, sw, sv, th, tr, uk, vi	None
Claude 3.7 Sonnet	en, es, pt, it, fr, id, de, ar, zh, ko, ja, hi, bn, sw, yo	None
QwQ-32B	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru
Kimi 1.5	en, zh	es, ar, hi, de, fr, pt, ja, ko, bn, sw, it, id, yo, my, th, ru

A.3 Datasets used to develop code-switched reasoning behavior taxonomy

Dataset	Type	Citation
s1K-1.1 and s1K-claude-3-7-sonnet	Coding, math, puzzle, science, legal, logic, humanities	Muennighoff et al. (2025b)
Global MMLU	Math, science, humanities, medical, business, legal	Singh et al. (2025)
MMLU ProX	Math, science, humanities, medical, business, legal	Xuan et al. (2025)
BIG-Bench Extra Hard	Puzzle, logic, commonsense, linguistics, spatial	Kazemi et al. (2025)
UniMoral	Cultural, social	Kumar & Jurgens (2025)
OpenMathReasoning	Math	Moshkov et al. (2025)
Kimi-K1.5-Distill-data	Math	Wang (2025); Ye et al. (2025)

A.4 Configuration for generating reasoning examples to develop taxonomy of code-switched reasoning behaviors

We generate reasoning examples on the BBEH, UniMoral, Global MMLU Lite, and MMLU ProX Lite datasets. All other data used in our experiments is drawn from preexisting reasoning example datasets.

BBEH Following the recommendations from the model creators,[3] for prompting DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Llama-8B we set `temperature=0.6,`

[2]The Gemma 3 paper and website do not specify which languages are supported, but we assume that the over 140 languages covered by the model include those languages that we prompt the model on in this work (Google, 2025b; Gemma Team et al., 2025).

[3]`https://huggingface.co/deepseek-ai/DeepSeek-R1`

top_p=0.95. We also set max_tokens=32768, which is the maximum allowed by the DeepSeek distilled models.

UniMoral For both Phi-4-mini-reasoning and DeepSeek-R1-Distill Llama-8B, we set temperature=0.6, top_p=0.95, and max_tokens=32768.

Global MMLU Lite For DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B, and DeepSeek-R1-Distill-Qwen-14B, we set temperature=0.6, top_p=0.9, and max_tokens=32768, again following the recommendations from the model creators.[4] For DeepSeek-R1-Distill-Qwen-32B and DeepSeek-R1-Distill-Llama-70B, we set temperature=1.0, top_p=1.0, and max_tokens=32768. For prompting DeepSeek-R1, we leave the API defaults as is and use the January 2025 model checkpoint.[5]

MMLU ProX Lite For Llama-3.1-Nemotron-Nano-8B-v1, we set temperature=0.6, top_p=0.95, and max_tokens=32768, following NVIDIA's recommendations.[6] We also turn reasoning mode on. For gemma-3-4b-it and DeepSeek-R1-Distill-Qwen-7B, we set temperature=0 and max_tokens=2048, as we initially followed the defaults in lm-evaluation-harness.[7] For DeepSeek-R1-Distill-Llama-8B, we set temperature=0 and max_tokens=32768.

[4] https://huggingface.co/deepseek-ai/DeepSeek-R1

[5] https://api-docs.deepseek.com/news/news250120

[6] https://build.nvidia.com/nvidia/llama-3_1-nemotron-nano-8b-v1/modelcard

[7] https://github.com/EleutherAI/lm-evaluation-harness/blob/main/lm_eval/tasks/mmlu_prox/en/_en_template_yaml

A.5 PROMPT FOR BRAINSTORMING CODE-SWITCHED REASONING BEHAVIOR TAXONOMY

Here is a problem and the reasoning process that a model generated when it tried to solve the problem.
Problem: (enclosed in double backticks)

``
{problem}
``

Reasoning process: (enclosed in triple backticks, the reasoning process has been split into distinct reasoning steps in the format of <step_idx><reasoning_step_content></step_idx>)

```
{reasoning}
```

You are tasked with analyzing any code-switching strategies used in the above reasoning, where code-switching is defined as the use of multiple scripts or languages. Your goal is to extract and describe patterns based on various criteria that characterize how the model uses code-switching in its reasoning.
Please follow these guidelines:
1. Identify multiple *meaningful criteria* that differentiate code-switching strategies. Each criterion should have a clear and descriptive name that reflects a real aspect of the code-switching process. **Do not use generic placeholders like 'Criterion 1'.**
2. For each criterion, provide a brief description.
3. Present your analysis in the following format, using <patterns> and </patterns> tags to enclose the list:

```
<patterns>
Descriptive Criterion Name (Description of the Criterion)
Descriptive Criterion Name (Description of the Criterion)
...
Descriptive Criterion Name (Description of the Criterion)
</patterns>
```

4. Do not include any additional explanations or commentary within the <patterns> tags.
5. If no code-switching is observed in the reasoning, simply return "<patterns></patterns>".

A.6 BERTOPIC MODELING DETAILS FOR CODE-SWITCHED REASONING BEHAVIOR TAXONOMY

All components of our topic modeling pipeline used in the development of the code-switched reasoning taxonomy are implemented in BERTopic. We use all-mpnet-base-v2 to embed the text of each criterion, UMAP for dimension reduction, HDBSCAN for clustering, and Gemini 2.5 Flash for generating topic representations. For UMAP dimension reduction, we set n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine'. For HDBSCAN clustering, we set min_cluster_size =15, metric='euclidean', cluster_selection_method ='eom', and prediction_data =True. We use CountVectorizer with stop_words=''english'' for topic tokenization and c-TF-IDF for extracting topic words. Finally, we use Gemini 2.5 Flash to generate more interpretable topic representations.

A.7 HUMAN EVALUATION OF LLM ANNOTATIONS

To validate the LLM-brainstormed criteria to differentiate code-switching strategies, we follow Lee et al. (2025) in asking our evaluator to answer the following yes-no question for each instance: "Are the automatically generated, detailed criteria plausible?"

A.8 WORD-LEVEL LANGUAGE IDENTIFICATION (LID)

For most languages we use whitespace to detect word boundaries. For Chinese, Japanese, Thai, and Burmese, which lack whitespace between words, we use jieba,[8], nagisa,[9], PyThaiNLP (Phatthiyaphaibun et al., 2024), and myanmar_words[10] to obtain word tokens.

For word-level language identification, we primarily rely on the *Lingua* library, which is computationally efficient, designed to work well on short texts, and covers most languages in our experiments, with the exception of Amharic, Burmese, and Igbo.[11] For Amharic and Burmese, which are written in non-Latin scripts, we use GlotScript to separate Amharic and Burmese words from English words (Kargaran et al., 2024). For Igbo, we use the FastText language identification model along with the English dictionary from Marchisio et al. (2024) to catch English words that are missed by FastText.

To validate our LID approach, we assemble a validation set of 140 reasoning examples generated from prompting Qwen3-Next-80B-A3B-Thinking in Amharic, Hindi, Igbo, Indonesian, Malay, Swahili, and Yoruba on Global MMLU (Singh et al., 2025). We achieve an average word-level LID accuracy of 0.933 on our validation set.

A.9 PROMPT FOR EVALUATING ACCURACY OF CODE-SWITCHING

Your task is to score the accuracy of the code-switching in the following text on a scale from 1 to 3, where 1 is the lowest and 3 is the highest. Accuracy measures whether the code-switched terms are used correctly.

Text to score:

```
{reasoning}
```

Use the following rubric to guide your scoring. The rubric is formatted as "<score>. <description>":

1. Low Accuracy: Code-switched terms are incorrect or inappropriate.

2. Moderate Accuracy: Most code-switched terms are appropriate but with minor mistakes.

3. High Accuracy: Code-switched terms are accurate and appropriately used.

Respond with only your score in the following format: <score>

[8] https://github.com/fxsjy/jieba
[9] https://github.com/taishi-i/nagisa
[10] https://github.com/linhtutkyawdev/myanmar_words
[11] https://github.com/pemistahl/lingua-py

A.10 PROMPT FOR EVALUATING FLUENCY OF CODE-SWITCHING

Your task is to score the fluency of the code-switching in the following text on a scale from 1 to 3, where 1 is the lowest and 3 is the highest. Fluency measures how natural and easy to understand the text is, considering grammar, syntax, and the smooth integration of code-switching.

Text to score:

`{reasoning}`

Use the following rubric to guide your scoring. The rubric is formatted as "`<score>. <description>`":

1. Low Fluency: The text is difficult to understand, awkward, or features poor grammar or syntax; or, the code-switching disrupts the flow of the text.

2. Moderate Fluency: The text is understandable but may have awkward or unnatural phrasing, grammar and syntax is acceptable, and code-switching is somewhat smooth but not perfectly integrated.

3. High Fluency: The text is natural and easy to understand; grammar and syntax are good; and code-switching is smooth and seamless, enhancing the sentence flow.

Respond with only your score in the following format: <score>

A.11 REASONING-RELATED SUBJECTS IN GLOBAL MMLU

- abstract_algebra
- business_ethics
- college_biology
- college_chemistry
- college_computer_science
- college_mathematics
- college_medicine
- college_physics
- computer_security
- conceptual_physics
- econometrics
- electrical_engineering
- elementary_mathematics
- formal_logic
- high_school_biology
- high_school_chemistry
- high_school_computer_science
- high_school_european_history
- high_school_mathematics
- high_school_physics
- high_school_statistics
- high_school_us_history

- high_school_world_history
- jurisprudence
- logical_fallacies
- machine_learning
- moral_scenarios
- professional_accounting
- professional_law
- professional_medicine
- professional_psychology

A.12 MT PIPELINE FOR SFT DATA SYNTHESIS

We use SeamlessM4T v2 for translation from English into all languages, except Malay (Barrault et al., 2023). For Malay, we use the 3.3B variant of NLLB-200, since unlike Seamless M4T v2, this model supports translation from English into Standard Malay (Costa-jussà et al., 2022). We select these models because they are the strongest available open-source massively multilingual MT models.

To further improve the quality of the translations, we use Chen et al.'s approach for translation refinement with the 70B Apertus model and filter the resulting translations for quality (Apertus et al., 2025).

Given a translation from the 3.3B variant of the NLLB-200 model or from the Seamless M4T v2 model, we prompt the 70B Apertus model with the following prompt sourced from Chen et al. to improve the translation:

```
Source: {English source text}
Bad translation: {Translation into target language}
Please give me a better {target language} translation without
    any explanation.
```

We use the default recommended sampling parameters for the Apertus model.[12]

Since Apertus does not always comply with the instructions not to provide additional explanation, we perform regex-based postprocessing to cleanup extraneous text. After postprocessing, we only keep the output from Apertus if the number of whitespace-separated tokens is less than twice as long as in the original translation. Otherwise, we retain the original machine translation from NLLB-200/Seamless M4T v2. We perform an additional quality filtering step to remove any translations that contain sequences of character or whitespace-separated token quadrigrams repeated more than 10 times, or sequences of the "#" or "*" characters repeated more than 10 times.

A.13 INSTANCE-LEVEL LANGUAGE IDENTIFICATION

Using the prompt below with the LID validation set from A.8, we obtain comparable accuracies of 0.892 for Qwen3-Next-80B-A3B-Instruct and 0.942 for Qwen3.5-27B (Yang et al., 2025).[13]

[12] https://huggingface.co/swiss-ai/Apertus-70B-2509
[13] https://qwen.ai/blog?id=qwen3.5

Identify the languages that appear in the following text:

`{text}`

Use the following language codes. The language codes are provided in the format "`<language_name>: <language_code>`":

`{language_codes}`

If you identify a language that is not listed above, please use the appropriate two-letter ISO 639 language code for that language.

Return your answer as a list in the following format: "`<language_code1>, <language_code2>, ...`"

If you are unable to identify any of the languages in the provided text, return "None"

A.14 SFT CONFIGURATION

We use the following configuration with LLaMA Factory for SFT (Zheng et al., 2024). We set the `cutoff_len` by computing the 95th percentile of the distribution of token lengths when considering the combined total length of input and target token sequences in each training dataset. `cutoff_len` is specific to each model, language, and SFT task. For the native-language reasoning, English-language reasoning, strategically code-switched reasoning, and synthetically code-switched reasoning tasks, we set `eval_steps` and `save_steps` to `10`. For SFT for machine translation and reasoning prompt translation into English, we set `eval_steps` and `save_steps` to `100`. We set `gradient_accumulation_steps` to ensure an effective batch size of `16`.

```
model_name_or_path: {deepseek-ai/DeepSeek-R1-Distill-Llama-8B,
    microsoft/Phi-4-reasoning, Qwen/Qwen3-8B}
stage: sft
do_train: true
finetuning_type: lora
lora_target: all
dataset_dir: {path2dataset_dir}
dataset: {dataset_name}
eval_dataset: {eval_dataset_name}
template: {deepseekr1, phi4, qwen3}
cutoff_len: {cutoff_len}
output_dir: {output_dir}
run_name: {run_name}
learning_rate: 1.0e-6
eval_strategy: steps
eval_steps: {eval_steps}
save_steps: {save_steps}
gradient_accumulation_steps: {grad_accum_steps}
overwrite_cache: true
preprocessing_num_workers: 16
logging_steps: 10
plot_loss: true
overwrite_output_dir: true
per_device_train_batch_size: 1
num_train_epochs: 3.0
lr_scheduler_type: cosine
warmup_ratio: 0.1
bf16: true
ddp_timeout: 180000000
per_device_eval_batch_size: 1
report_to: wandb
```

A.15 GLOBAL PIQA NATIVE PROMPTS

A.16 ACCURACY ON GLOBAL PIQA AND CORE CORPUS SFT TEST SET (CURATED FROM GLOBAL MMLU)

Language codes: hi = Hindi, id = Indonesian, ms = Malay, am = Amharic, ig = Igbo, sw = Swahili, yo = Yoruba.

SFT conditions: No-SFT baseline = reasoning model without any SFT intervention applied; English reasoning = fine-tune for reasoning in English; Machine translation = fine-tune to perform machine translation; Native reasoning = fine-tune to perform native-language reasoning; Strategic/Synthetic CSW = fine-tune to perform strategically/synthetically code-switched reasoning; Translated prompts = fine-tune to translate reasoning prompts into English. Yoruba has no native-reasoning or synthetic-CSW fine-tuning conditions due to insufficient synthesized data.

Evaluation configurations: CoRe Global MMLU test set = CoRe corpus test set curated from Global MMLU questions; Global PIQA = Global PIQA using the original English-framed prompts from Chang et al. (2026); Global PIQA native = Global PIQA using fully native-language prompts.

Table 4: Accuracy by condition, language, model, and evaluation configuration. Each cell reports the point estimate with a Wilson 95% confidence interval as a subscript.

Condition	hi	id	ms	am	ig	sw	yo
CoRe Global MMLU test set — DeepSeek-R1-Distill-Llama-8B ($n = 500$ per cell)							
No-SFT baseline	$0.46_{[.41,.50]}$	$0.53_{[.49,.58]}$	$0.53_{[.48,.57]}$	$0.26_{[.22,.30]}$	$0.25_{[.22,.29]}$	$0.16_{[.13,.19]}$	$0.29_{[.25,.33]}$
English reasoning	$0.45_{[.41,.49]}$	$0.59_{[.55,.63]}$	$0.52_{[.48,.57]}$	$0.26_{[.22,.30]}$	$0.25_{[.22,.29]}$	$0.19_{[.16,.23]}$	$0.30_{[.26,.34]}$

Condition	hi	id	ms	am	ig	sw	yo
Machine translation	$0.42_{[.38,.47]}$	$0.55_{[.51,.60]}$	$0.52_{[.48,.56]}$	$0.25_{[.21,.29]}$	$0.30_{[.26,.34]}$	$0.23_{[.20,.27]}$	$0.26_{[.22,.30]}$
Native reasoning	$0.44_{[.40,.48]}$	$0.56_{[.52,.61]}$	$0.56_{[.51,.60]}$	$0.25_{[.21,.29]}$	$0.29_{[.25,.33]}$	$0.22_{[.19,.26]}$	–
Strategic CSW	$0.44_{[.39,.48]}$	$0.58_{[.53,.62]}$	$0.55_{[.51,.60]}$	$0.25_{[.22,.29]}$	$0.25_{[.21,.29]}$	$0.21_{[.17,.24]}$	$0.28_{[.24,.32]}$
Synthetic CSW	$0.45_{[.41,.50]}$	$0.58_{[.53,.62]}$	$0.55_{[.51,.60]}$	$0.27_{[.23,.31]}$	$0.27_{[.23,.31]}$	$0.19_{[.16,.23]}$	–
Translated prompts	$0.46_{[.41,.50]}$	$0.55_{[.50,.59]}$	$0.56_{[.51,.60]}$	$0.27_{[.23,.31]}$	$0.30_{[.27,.35]}$	$0.24_{[.20,.28]}$	$0.26_{[.22,.30]}$
CoRe Global MMLU test set — Phi-4-reasoning ($n = 500$ per cell)							
No-SFT baseline	$0.68_{[.64,.72]}$	$0.72_{[.67,.75]}$	$0.70_{[.65,.73]}$	$0.26_{[.22,.30]}$	$0.33_{[.29,.37]}$	$0.61_{[.56,.65]}$	$0.31_{[.27,.35]}$
English reasoning	$0.67_{[.63,.71]}$	$0.72_{[.68,.76]}$	$0.68_{[.64,.72]}$	$0.23_{[.20,.27]}$	$0.29_{[.25,.33]}$	$0.57_{[.53,.61]}$	$0.27_{[.23,.31]}$
Machine translation	$0.67_{[.63,.71]}$	$0.69_{[.65,.73]}$	$0.69_{[.65,.73]}$	$0.25_{[.21,.29]}$	$0.29_{[.25,.33]}$	$0.61_{[.56,.65]}$	$0.26_{[.22,.30]}$
Native reasoning	$0.68_{[.64,.72]}$	$0.72_{[.68,.76]}$	$0.67_{[.63,.71]}$	$0.24_{[.21,.28]}$	$0.31_{[.27,.35]}$	$0.60_{[.55,.64]}$	–
Strategic CSW	$0.68_{[.64,.72]}$	$0.68_{[.64,.72]}$	$0.67_{[.63,.71]}$	$0.21_{[.18,.25]}$	$0.30_{[.26,.34]}$	$0.60_{[.55,.64]}$	$0.27_{[.23,.31]}$
Synthetic CSW	$0.70_{[.66,.74]}$	$0.70_{[.66,.74]}$	$0.67_{[.63,.71]}$	$0.24_{[.21,.28]}$	$0.31_{[.27,.35]}$	$0.64_{[.60,.68]}$	–
Translated prompts	$0.68_{[.64,.72]}$	$0.73_{[.69,.77]}$	$0.68_{[.64,.72]}$	$0.24_{[.21,.28]}$	$0.29_{[.25,.33]}$	$0.62_{[.57,.66]}$	$0.25_{[.22,.29]}$
CoRe Global MMLU test set — Qwen3-8B ($n = 500$ per cell)							
No-SFT baseline	$0.74_{[.70,.77]}$	$0.79_{[.76,.83]}$	$0.76_{[.72,.80]}$	$0.31_{[.27,.35]}$	$0.31_{[.27,.35]}$	$0.41_{[.37,.45]}$	$0.28_{[.24,.32]}$
English reasoning	$0.73_{[.69,.77]}$	$0.77_{[.73,.80]}$	$0.78_{[.74,.81]}$	$0.31_{[.27,.35]}$	$0.34_{[.30,.38]}$	$0.41_{[.36,.45]}$	$0.27_{[.23,.31]}$
Machine translation	$0.74_{[.70,.77]}$	$0.77_{[.73,.80]}$	$0.73_{[.69,.77]}$	$0.39_{[.35,.43]}$	$0.39_{[.34,.43]}$	$0.48_{[.43,.52]}$	$0.31_{[.27,.35]}$
Native reasoning	$0.75_{[.71,.79]}$	$0.78_{[.74,.82]}$	$0.74_{[.70,.78]}$	$0.30_{[.26,.34]}$	$0.33_{[.29,.37]}$	$0.44_{[.40,.49]}$	–
Strategic CSW	$0.74_{[.70,.78]}$	$0.78_{[.74,.81]}$	$0.77_{[.73,.80]}$	$0.32_{[.28,.36]}$	$0.32_{[.28,.37]}$	$0.45_{[.40,.49]}$	$0.32_{[.28,.36]}$
Synthetic CSW	$0.74_{[.70,.78]}$	$0.77_{[.73,.81]}$	$0.76_{[.72,.80]}$	$0.30_{[.26,.34]}$	$0.31_{[.27,.35]}$	$0.43_{[.38,.47]}$	–
Translated prompts	$0.73_{[.69,.77]}$	$0.76_{[.72,.79]}$	$0.77_{[.73,.80]}$	$0.32_{[.28,.37]}$	$0.31_{[.27,.35]}$	$0.37_{[.33,.42]}$	$0.29_{[.25,.33]}$
Global PIQA — DeepSeek-R1-Distill-Llama-8B ($n = 100$ per cell)							
No-SFT baseline	$0.80_{[.71,.87]}$	$0.87_{[.79,.92]}$	$0.81_{[.72,.87]}$	$0.50_{[.40,.60]}$	$0.54_{[.44,.63]}$	$0.76_{[.67,.83]}$	$0.52_{[.42,.62]}$
English reasoning	$0.80_{[.71,.87]}$	$0.78_{[.69,.85]}$	$0.80_{[.71,.87]}$	$0.48_{[.38,.58]}$	$0.55_{[.45,.64]}$	$0.61_{[.51,.70]}$	$0.50_{[.40,.60]}$
Machine translation	$0.72_{[.63,.80]}$	$0.81_{[.72,.87]}$	$0.73_{[.64,.81]}$	$0.41_{[.32,.51]}$	$0.53_{[.43,.62]}$	$0.66_{[.56,.75]}$	$0.47_{[.38,.57]}$
Native reasoning	$0.77_{[.68,.84]}$	$0.82_{[.73,.88]}$	$0.73_{[.64,.81]}$	$0.42_{[.33,.52]}$	$0.57_{[.47,.66]}$	$0.59_{[.49,.68]}$	–
Strategic CSW	$0.74_{[.65,.82]}$	$0.78_{[.69,.85]}$	$0.74_{[.65,.82]}$	$0.47_{[.38,.57]}$	$0.54_{[.44,.63]}$	$0.68_{[.58,.76]}$	$0.62_{[.52,.71]}$
Synthetic CSW	$0.81_{[.72,.87]}$	$0.80_{[.71,.87]}$	$0.73_{[.64,.81]}$	$0.50_{[.40,.60]}$	$0.60_{[.50,.69]}$	$0.60_{[.50,.69]}$	–
Translated prompts	$0.76_{[.67,.83]}$	$0.80_{[.71,.87]}$	$0.78_{[.69,.85]}$	$0.41_{[.32,.51]}$	$0.49_{[.39,.59]}$	$0.56_{[.46,.65]}$	$0.51_{[.41,.61]}$
Global PIQA — Phi-4-reasoning ($n = 100$ per cell)							
No-SFT baseline	$0.90_{[.83,.94]}$	$0.94_{[.88,.97]}$	$0.86_{[.78,.91]}$	$0.67_{[.57,.75]}$	$0.56_{[.46,.65]}$	$0.79_{[.70,.86]}$	$0.49_{[.39,.59]}$
English reasoning	$0.96_{[.90,.98]}$	$0.93_{[.86,.97]}$	$0.87_{[.79,.92]}$	$0.57_{[.47,.66]}$	$0.58_{[.48,.67]}$	$0.81_{[.72,.87]}$	$0.61_{[.51,.70]}$
Machine translation	$0.93_{[.86,.97]}$	$0.93_{[.86,.97]}$	$0.86_{[.78,.91]}$	$0.57_{[.47,.66]}$	$0.58_{[.48,.67]}$	$0.81_{[.72,.87]}$	$0.61_{[.51,.70]}$
Native reasoning	$0.94_{[.88,.97]}$	$0.93_{[.86,.97]}$	$0.89_{[.81,.94]}$	$0.57_{[.47,.66]}$	$0.66_{[.56,.75]}$	$0.82_{[.73,.88]}$	–
Strategic CSW	$0.92_{[.85,.96]}$	$0.94_{[.88,.97]}$	$0.84_{[.76,.90]}$	$0.67_{[.57,.75]}$	$0.56_{[.46,.65]}$	$0.85_{[.77,.91]}$	$0.63_{[.53,.72]}$
Synthetic CSW	$0.93_{[.86,.97]}$	$0.94_{[.88,.97]}$	$0.85_{[.77,.91]}$	$0.72_{[.63,.80]}$	$0.51_{[.41,.61]}$	$0.82_{[.73,.88]}$	–
Translated prompts	$0.94_{[.88,.97]}$	$0.95_{[.89,.98]}$	$0.86_{[.78,.91]}$	$0.61_{[.51,.70]}$	$0.55_{[.45,.64]}$	$0.82_{[.73,.88]}$	$0.53_{[.43,.62]}$
Global PIQA — Qwen3-8B ($n = 100$ per cell)							
No-SFT baseline	$0.87_{[.79,.92]}$	$0.89_{[.81,.94]}$	$0.80_{[.71,.87]}$	$0.61_{[.51,.70]}$	$0.53_{[.43,.62]}$	$0.72_{[.63,.80]}$	$0.46_{[.37,.56]}$
English reasoning	$0.87_{[.79,.92]}$	$0.88_{[.80,.93]}$	$0.83_{[.74,.89]}$	$0.62_{[.52,.71]}$	$0.59_{[.49,.68]}$	$0.73_{[.64,.81]}$	$0.44_{[.35,.54]}$
Machine translation	$0.83_{[.74,.89]}$	$0.89_{[.81,.94]}$	$0.78_{[.69,.85]}$	$0.68_{[.58,.76]}$	$0.55_{[.45,.64]}$	$0.67_{[.57,.75]}$	$0.45_{[.36,.55]}$
Native reasoning	$0.89_{[.81,.94]}$	$0.89_{[.81,.94]}$	$0.78_{[.69,.85]}$	$0.68_{[.58,.76]}$	$0.61_{[.51,.70]}$	$0.71_{[.61,.79]}$	–
Strategic CSW	$0.84_{[.76,.90]}$	$0.88_{[.80,.93]}$	$0.83_{[.74,.89]}$	$0.62_{[.52,.71]}$	$0.54_{[.44,.63]}$	$0.71_{[.61,.79]}$	$0.50_{[.40,.60]}$
Synthetic CSW	$0.88_{[.80,.93]}$	$0.92_{[.85,.96]}$	$0.77_{[.68,.84]}$	$0.61_{[.51,.70]}$	$0.56_{[.46,.65]}$	$0.61_{[.51,.70]}$	–
Translated prompts	$0.86_{[.78,.91]}$	$0.88_{[.80,.93]}$	$0.82_{[.73,.88]}$	$0.70_{[.60,.78]}$	$0.62_{[.52,.71]}$	$0.72_{[.63,.80]}$	$0.59_{[.49,.68]}$
Global PIQA native — DeepSeek-R1-Distill-Llama-8B ($n = 100$ per cell)							
No-SFT baseline	$0.80_{[.71,.87]}$	$0.84_{[.76,.90]}$	$0.73_{[.64,.81]}$	$0.50_{[.40,.60]}$	$0.16_{[.10,.24]}$	$0.49_{[.39,.59]}$	$0.50_{[.40,.60]}$
English reasoning	$0.71_{[.61,.79]}$	$0.80_{[.71,.87]}$	$0.70_{[.60,.78]}$	$0.40_{[.31,.50]}$	$0.24_{[.17,.33]}$	$0.44_{[.35,.54]}$	$0.48_{[.38,.58]}$
Machine translation	$0.54_{[.44,.63]}$	$0.78_{[.69,.85]}$	$0.73_{[.64,.81]}$	$0.42_{[.33,.52]}$	$0.46_{[.37,.56]}$	$0.43_{[.34,.53]}$	$0.36_{[.27,.46]}$
Native reasoning	$0.68_{[.58,.76]}$	$0.80_{[.71,.87]}$	$0.72_{[.63,.80]}$	$0.43_{[.34,.53]}$	$0.23_{[.16,.32]}$	$0.47_{[.38,.57]}$	–
Strategic CSW	$0.79_{[.70,.86]}$	$0.77_{[.68,.84]}$	$0.76_{[.67,.83]}$	$0.39_{[.30,.49]}$	$0.24_{[.17,.33]}$	$0.52_{[.42,.62]}$	$0.46_{[.37,.56]}$
Synthetic CSW	$0.75_{[.66,.82]}$	$0.80_{[.71,.87]}$	$0.76_{[.67,.83]}$	$0.44_{[.35,.54]}$	$0.25_{[.18,.34]}$	$0.51_{[.41,.61]}$	–
Translated prompts	$0.77_{[.68,.84]}$	$0.82_{[.73,.88]}$	$0.70_{[.60,.78]}$	$0.35_{[.26,.45]}$	$0.31_{[.23,.41]}$	$0.43_{[.34,.53]}$	$0.42_{[.33,.52]}$
Global PIQA native — Phi-4-reasoning ($n = 100$ per cell)							
No-SFT baseline	$0.94_{[.88,.97]}$	$0.95_{[.89,.98]}$	$0.83_{[.74,.89]}$	$0.52_{[.42,.62]}$	$0.52_{[.42,.62]}$	$0.81_{[.72,.87]}$	$0.50_{[.40,.60]}$
English reasoning	$0.93_{[.86,.97]}$	$0.95_{[.89,.98]}$	$0.86_{[.78,.91]}$	$0.54_{[.44,.63]}$	$0.53_{[.43,.62]}$	$0.83_{[.74,.89]}$	$0.49_{[.39,.59]}$

Condition	hi	id	ms	am	ig	sw	yo
Machine translation	$0.94_{[.88,.97]}$	$0.95_{[.89,.98]}$	$0.84_{[.76,.90]}$	$0.54_{[.44,.63]}$	$0.68_{[.58,.76]}$	$0.83_{[.74,.89]}$	$0.54_{[.44,.63]}$
Native reasoning	$0.96_{[.90,.98]}$	$0.93_{[.86,.97]}$	$0.90_{[.83,.94]}$	$0.58_{[.48,.67]}$	$0.63_{[.53,.72]}$	$0.84_{[.76,.90]}$	–
Strategic CSW	$0.95_{[.89,.98]}$	$0.94_{[.88,.97]}$	$0.84_{[.76,.90]}$	$0.58_{[.48,.67]}$	$0.63_{[.53,.72]}$	$0.83_{[.74,.89]}$	$0.59_{[.49,.68]}$
Synthetic CSW	$0.94_{[.88,.97]}$	$0.95_{[.89,.98]}$	$0.89_{[.81,.94]}$	$0.46_{[.37,.56]}$	$0.53_{[.43,.62]}$	$0.84_{[.76,.90]}$	–
Translated prompts	$0.94_{[.88,.97]}$	$0.96_{[.90,.98]}$	$0.81_{[.72,.87]}$	$0.54_{[.44,.63]}$	$0.59_{[.49,.68]}$	$0.82_{[.73,.88]}$	$0.53_{[.43,.62]}$
Global PIQA native — Qwen3-8B ($n = 100$ per cell)							
No-SFT baseline	$0.85_{[.77,.91]}$	$0.90_{[.83,.94]}$	$0.84_{[.76,.90]}$	$0.63_{[.53,.72]}$	$0.56_{[.46,.65]}$	$0.73_{[.64,.81]}$	$0.48_{[.38,.58]}$
English reasoning	$0.86_{[.78,.91]}$	$0.89_{[.81,.94]}$	$0.80_{[.71,.87]}$	$0.58_{[.48,.67]}$	$0.56_{[.46,.65]}$	$0.66_{[.56,.75]}$	$0.53_{[.43,.62]}$
Machine translation	$0.70_{[.60,.78]}$	$0.89_{[.81,.94]}$	$0.83_{[.74,.89]}$	$0.58_{[.48,.67]}$	$0.51_{[.41,.61]}$	$0.67_{[.57,.75]}$	$0.32_{[.24,.42]}$
Native reasoning	$0.87_{[.79,.92]}$	$0.89_{[.81,.94]}$	$0.79_{[.70,.86]}$	$0.60_{[.50,.69]}$	$0.49_{[.39,.59]}$	$0.72_{[.63,.80]}$	–
Strategic CSW	$0.87_{[.79,.92]}$	$0.89_{[.81,.94]}$	$0.82_{[.73,.88]}$	$0.59_{[.49,.68]}$	$0.56_{[.46,.65]}$	$0.71_{[.61,.79]}$	$0.53_{[.43,.62]}$
Synthetic CSW	$0.86_{[.78,.91]}$	$0.90_{[.83,.94]}$	$0.85_{[.77,.91]}$	$0.60_{[.50,.69]}$	$0.53_{[.43,.62]}$	$0.75_{[.66,.82]}$	–
Translated prompts	$0.89_{[.81,.94]}$	$0.89_{[.81,.94]}$	$0.79_{[.70,.86]}$	$0.62_{[.52,.71]}$	$0.50_{[.40,.60]}$	$0.74_{[.65,.82]}$	$0.45_{[.36,.55]}$

A.17 Accuracy on English subsets of Global PIQA and CoRe corpus SFT test set (curated from Global MMLU)

English scores for the No-SFT baseline (no fine-tuned condition has been evaluated on English).

Table 5: Accuracy for English, No-SFT baseline only. Each cell reports the point estimate with a Wilson 95% confidence interval as a subscript; see appendix A.16 above for the evaluation configuration legend.

Model	CoRe Global MMLU test set	Global PIQA
DeepSeek-R1-Distill-Llama-8B	$0.69_{[.65,.73]}$	$0.76_{[.67,.83]}$
Phi-4-reasoning	$0.75_{[.71,.79]}$	$0.89_{[.81,.94]}$
Qwen3-8B	$0.83_{[.80,.86]}$	$0.89_{[.81,.94]}$

A.18 Supervised fine-tuning task dataset synthesis details

Native-language reasoning. This task serves as a monolingual baseline against which to compare fine-tuning tasks which use examples of code-switching. Given a prompt p_l in language l, the model is fine-tuned to produce its reasoning r_l and final answer a_l in the same language l, yielding (p_l, r_l, a_l). Given some p_l, we synthesize r_l and a_l by prompting Qwen3-Next-80B-A3B-Thinking with the equivalent English prompt p_e from Global MMLU (see section 4.1). We filter for instances with correct answers using the reference answers in Global MMLU. To ensure efficient reasoning, we then filter correct reasoning demonstrations by length (in number of output tokens), using a model- and language-specific cutoff.[14] Finally, we translate the English reasoning r_e and answers a_e into the target language to yield r_l and a_l. We source the prompt p_l directly from Global MMLU.

Machine translation into English. We observed that reasoning models often code-switch for the purpose of translation into a higher-resource language, e.g., English (section 3). Therefore, improving these models' translation abilities might help them code-switch to solve reasoning problems more effectively. To test this hypothesis, we fine-tune our models to perform translation from each of the languages in section 4.1 into English, sourcing examples from NLLB. The fine-tuning instances for this task consist of a prompt p_l instructing the model to translate an example non-English sentence into English, followed by empty reasoning r and the English translation of the sentence as the final answer a_e: (p_l, r, a_e).

Reasoning prompt translation into English. Similar to the motivation for the general machine translation task, this task targets models' ability to translate reasoning prompts—rather than the reasoning itself—into English. We use non-English prompts and their English translations from Global MMLU (see section 4.1) as examples for fine-tuning. More formally, the fine-tuning instances for

[14] We compute the 95th percentile of output lengths (in tokens) for correct reasoning instances from each baseline (non–fine-tuned) reasoning model on the validation set in each language.

this task consist of a prompt p_l instructing the model to translate an example non-English reasoning prompt into English, followed by empty reasoning r and the English translation of the prompt as the final answer a_e: (p_l, r, a_e).

English-language reasoning. Given a prompt p_l in language l, the model is fine-tuned to generate English-language reasoning r_e before returning its final answer a_l in the same language l as the prompt: (p_l, r_e, a_l). This task is motivated by the **Form** dimension of our taxonomy in section 3, which captures that models may switch into a different language for the entire reasoning process (as represented by this task), or interleave multiple languages during reasoning (as represented by the following tasks).

Strategically code-switched reasoning. Testing Li et al.'s view that code-switching is a strategic reasoning behavior, this condition fine-tunes the student reasoning models on examples of code-switched reasoning from Qwen3-Next-80B-A3B-Thinking. This task is designed assuming that as a strong multilingual reasoner, Qwen3-Next-80B-A3B-Thinking can generate examples of "good" code-switches for reasoning. Given prompt p_l in language l, we generate candidate code-switched reasoning demonstrations r_{csw} and final answers a_l in language l by prompting Qwen3-Next-80B-A3B-Thinking with p_l. Next, we filter for instances where multiple languages are present in r_{csw}. We detect code-switched reasoning instances using Qwen3-Next-80B-A3B-Instruct and the prompt from appendix A.13 (Yang et al., 2025). This yields fine-tuning instance (p_l, r_{csw}, a_l).

Synthetically code-switched reasoning. Given prompt p_l in language l, this task also fine-tunes the student reasoning models on examples of code-switched reasoning r_{csw} to produce final answers a_l in language l. For p_l in language l, we generate reasoning and answers in both English and l (r_e, a_e, r_l, a_l) using the same approach as for the *native-language reasoning* task. We use newlines to segment the reasoning into steps as done in Chen et al. (2025), then randomly select half of these steps to be in English, with the other half remaining in language l. We splice together the final code-switched reasoning trace r_{csw} from r_e and r_l, yielding fine-tuning instance (p_l, r_{csw}, a_l). The key contrast between this task and the *strategically code-switched reasoning* task above is that the code-switches are randomly generated, allowing us to confirm the importance of **coherent** code-switching (one of our main taxonomy dimensions in section 3) for overall performance. We expect that the random code-switching in this task is less coherent than the *strategically code-switched reasoning*.

A.19 Full forest plots (pooled, Global MMLU, and Global PIQA) for association between code-switching and final answer accuracy

Figure 5 shows the relationship between code-switching behaviors and final answer correctness for evaluations on Global MMLU, Global PIQA, and both datasets pooled.

A.20 Full heatmaps (pooled, Global MMLU, and Global PIQA)

Figure 6 shows the effect of our SFT interventions on the Global MMLU and Global PIQA datasets separately, in addition to the pooled view shown in the main paper text.

A.21 Gemma teacher generalization experiments

Figures 7 and 8 show that our framework generalizes across teacher models.

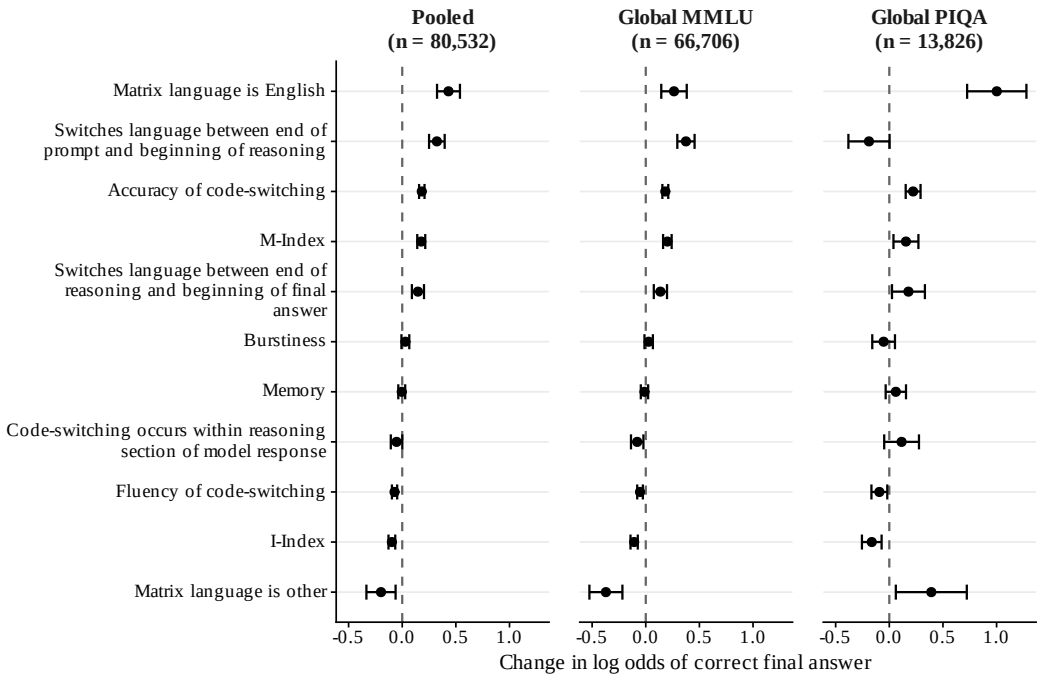

Figure 5: Across all evaluated models, languages, and datasets, certain code-switching behaviors in reasoning are associated with higher odds of a correct final answer. In particular, featuring English as the matrix language, switching into a different language than the target prompt language when transitioning from the prompt to the beginning of reasoning, finishing the reasoning in a different language before switching back to the target language for the final answer, higher accuracy of code-switching, and higher M-Index are all significantly associated with better reasoning performance.

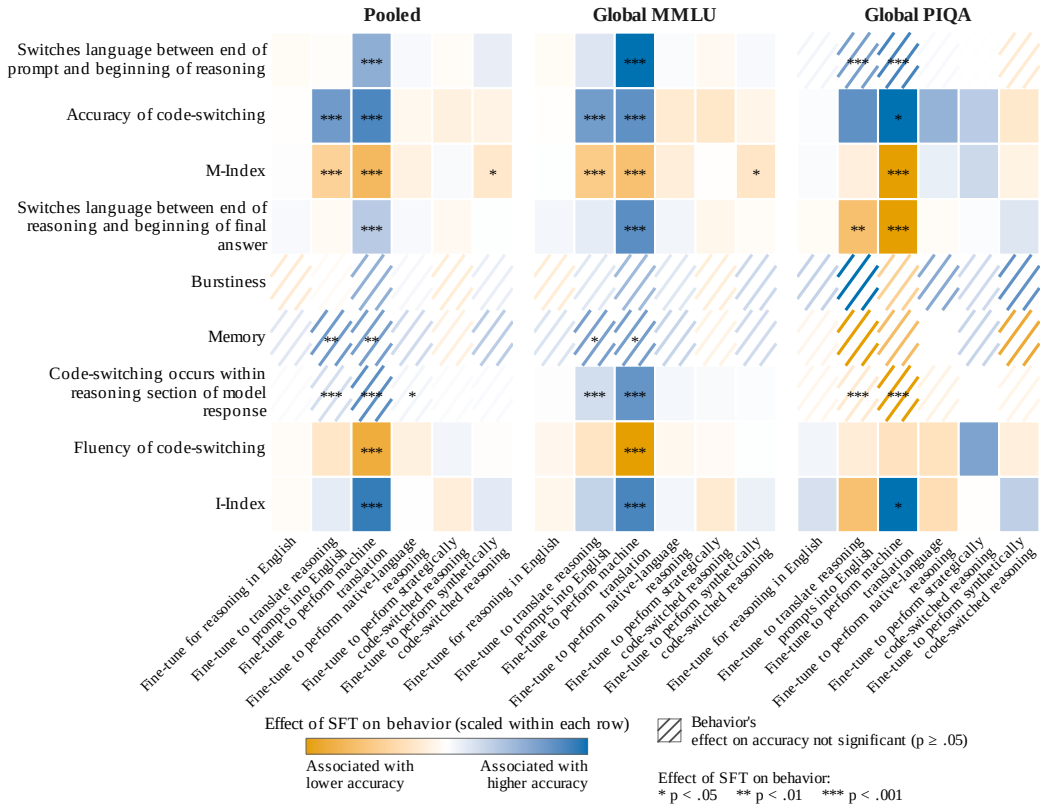

Figure 6: Across all models, languages, and datasets evaluated, our translation-based SFT intervention tasks (translating reasoning prompts into English and performing general machine translation into English) emerged as useful approaches for promoting code-switched reasoning behaviors associated with higher final answer accuracy.

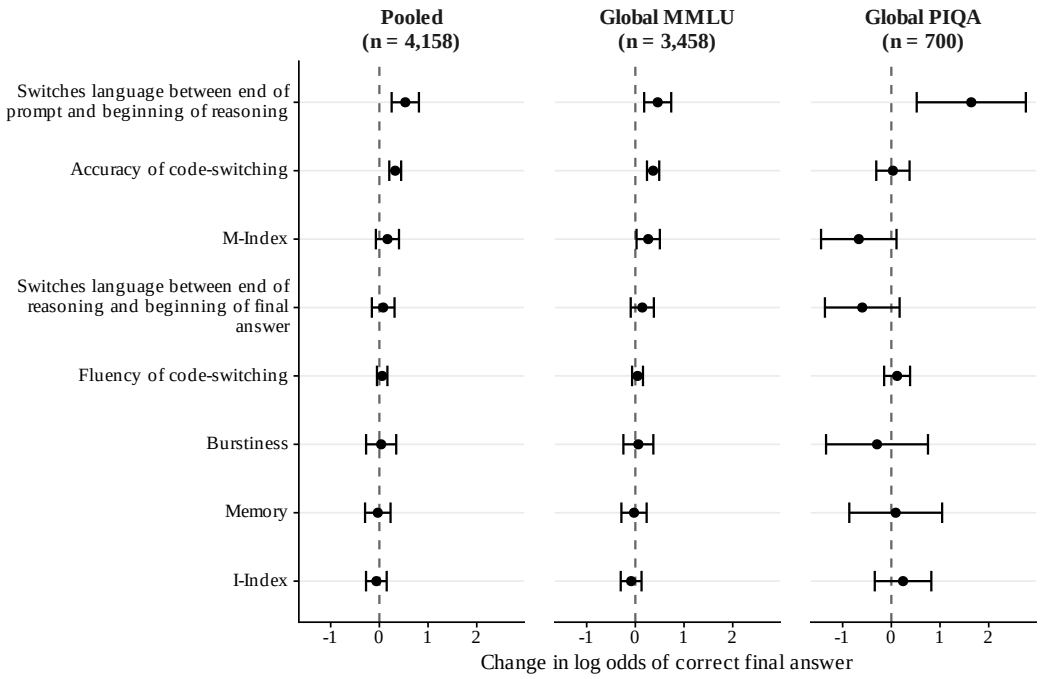

Figure 7: In addition to evaluating models fine-tuned on reasoning demonstrations from Qwen3-Next-80B-A3B-Thinking, our results also show that our approach generalizes to other teacher models (here, gemma-4-26B-A4B-it). We see that in the reasoning of models fine-tuned on gemma-4-26B-A4B-it reasoning demonstrations, certain key code-switched reasoning behaviors (e.g., accuracy of code-switching) remain associated with better performance, in line with our results from the Qwen3-Next-80B-A3B-Thinking teacher experiments.

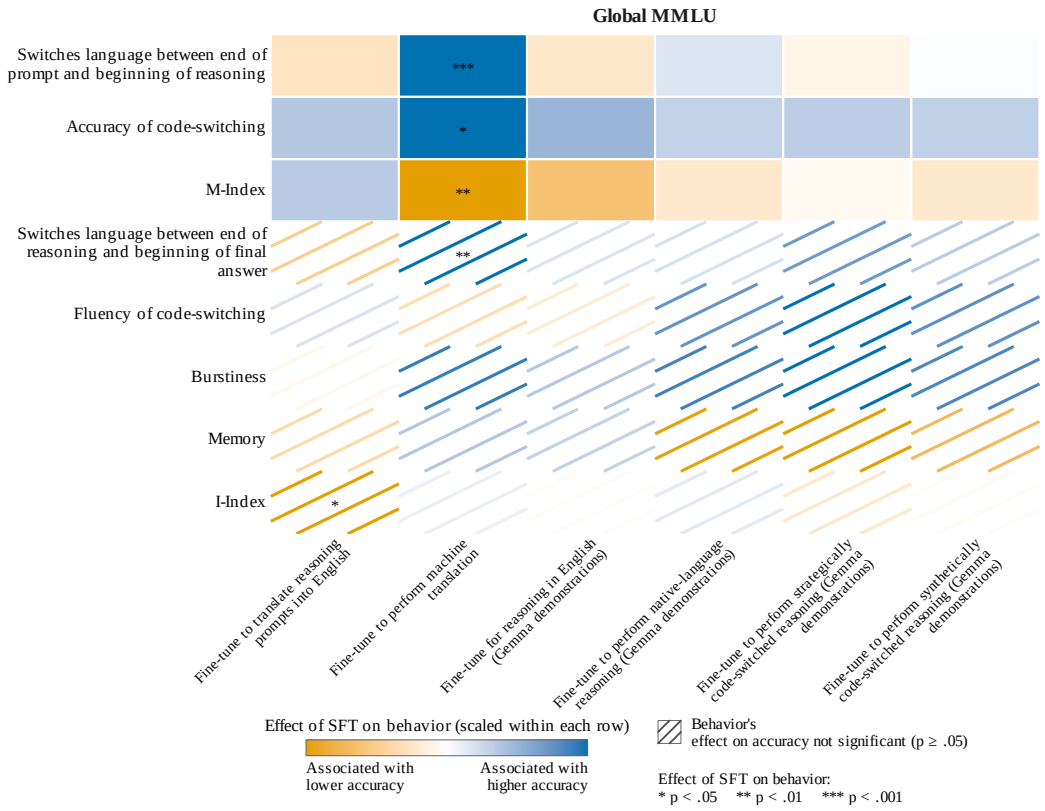

Figure 8: When using gemma-4-26B-A4B-it as the teacher model instead of Qwen3-Next-80B-A3B-Thinking, we see that our SFT interventions still significantly boost certain helpful code-switched reasoning behaviors.