# Are Multilingual Models Actually Improving? Isolating True Cross-Lingual Transfer

Anonymous ACL submission

## Abstract

Cross-lingual transfer is a model's ability to generalize capabilities from well-represented source languages to under-represented target languages. Existing measures of a model's transfer strength conflate improvements in transfer with general improvements to accuracy in the source language. We advocate for an alternate metric that reliably captures transfer strength called Hardness Adjusted Transfer (HAT) Score, and use it to derive multiple insights on factors influencing transfer strength. Our analysis across twenty diverse language models and three popular mainstream multilingual benchmarks argues that 1) transfer in small models is not broken, 2) we are making slower than expected progress in cross-lingual transfer with model size, and 3) we have made clear progress over time.

## 1 Introduction

Most large language models (LLMs) officially support multiple non-English languages (Comanici et al., 2025; Singh et al., 2026; Anthropic, 2025). However, significant performance gaps across languages are known to exist (Cohere Labs, 2024; Hu et al., 2025b), thus putting certain demographics at a technological disadvantage. Cross-lingual transfer (XLT), the ability of an LLM to transfer performance from a source language (such as English) to other target languages, is critical for achieving cross-lingual performance parity (Pires et al., 2019; Hu et al., 2020; Philippy et al., 2023). Improvements on multilingual benchmarks are commonly interpreted as evidence of improved XLT (Hu et al., 2020). In this work, we examine this assumption carefully and question whether gains in target language performance necessarily signal improved XLT.

Transfer strength is characterized by the fraction of performance translating from source to target languages. Therefore, we may declare improvements in transfer if target accuracy improved for the same value of source accuracy. In practice, however, source and target accuracies are linearly coupled with goodness of fit ($R^2$) exceeding 0.94 as shown in Figure 1, i.e., target accuracy follows predictably from source accuracy. The surprising linear trend is further discussed in Section 3. The coupled performance growth calls to question the standard measures of XLT such as (1) the absolute performance on target language, or (2) relative performance with source language, because they do not adjust for general improvements to source accuracy.

Both source and target accuracies may increase with general improvements such as factuality, reasoning, or model size, but increased target accuracy, a typical measure of cross-lingual transfer, need not imply progress on transfer. Source-target performance gap is another standard measure of transfer. We observe from the left plot of Figure 1 that both target accuracy and performance gap increase as source accuracy increases. While increasing target accuracy imply progress on XLT, increasing gap indicates worsening XLT. Together, they paint a conflicting narrative on XLT progress, which warrants a closer study.

We argue that an ideal transfer metric should measure improvements in expected target accuracies while adjusting for improvements in source accuracies. More precisely, in order to measure XLT, it is important to disentangle the effect of target improvements that simply follow from source improvements and measure when the model's accuracy lies above the expected correlation line (shown in Figure 1).

Motivated by these observations, this paper rethinks how we define progress in cross-lingual transfer. Our contributions are summarized below.

- We show that existing measures of cross-lingual gaps do not reflect true progress on

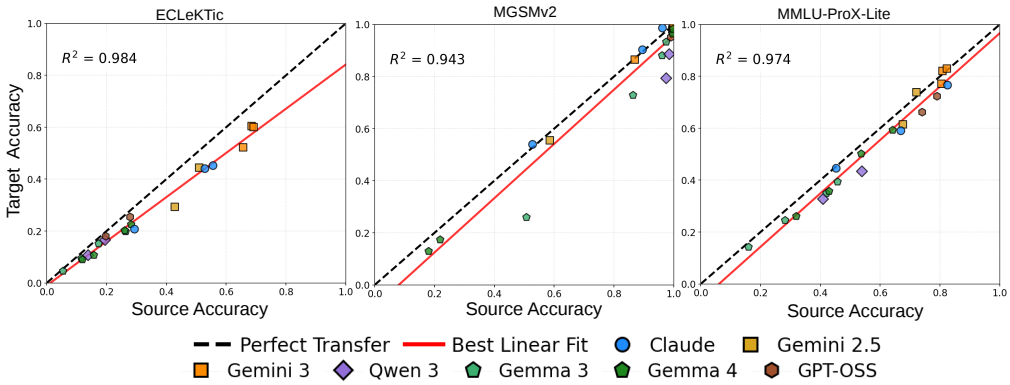



Figure 1: **Target accuracy scales predictably from source accuracy.** Evaluation with twenty diverse LLMs across multilingual benchmarks ECLeKTic, MGSMv2, and MMLU-ProX-Lite reveals a linear trend. For MGSMv2 and MMLU-ProX-Lite, we set English as source language, while for ECLeKTic, the source language varies per question as defined by the dataset. Target is a subset of non-source languages. Because the improvements in source and target accuracies are coupled, traditional XLT metrics conflate general source-language improvements with transfer. Please see Section 3 for more details and context.

XLT, and derive an ideal measure from first principles.

- Leveraging our calibrated XLT measure, we present insights on XLT influencing model choices such as thinking, parameter counts and language scripts.
- We show a linear trend between source and target accuracies across twenty models and three multilingual datasets spanning factuality and reasoning. The strong linear trend is interesting given the diversity in model class, size and thinking budget. We analyze progress over time and find that MMLU-ProX-Lite and MGSMv2 are both nearly solved in terms of XLT.

## 2 Related Work

**Multilingual Generalization.** Recent LLM-era benchmarks, including XTREME (Hu et al., 2020), XSQuAD (Artetxe et al., 2020), MGSM (Shi et al., 2022), MMLU-ProX (Xuan et al., 2025), ECLeKTic (Goldman et al., 2025), and MuBench (Han et al., 2025), shift multilingual evaluation towards assessing generative capabilities such as factual question answering and reasoning. This broader evaluation landscape revealed inconsistencies in model performance across languages of different resource, script (Xu et al., 2025; Hengle et al., 2024; Han et al., 2025; Xuan et al., 2025). Complementary analyses suggest that these failures reflect differences in internal alignment, shared multilingual computation, and, in some cases, English-mediated latent processing (Zhao et al., 2024; Ifergan et al., 2024; Tang et al., 2024).

**Metrics for XLT.** Typically, XLT is measured via target accuracy (Xuan et al., 2025; Hu et al., 2020) or the cross-lingual transfer gap (XLT gap) (Hu et al., 2020; Chi et al., 2021), i.e., the absolute difference between source and target accuracy. Recently, ECLeKTic used a "Transfer score" that measures target accuracy conditioned on correct source prediction (Goldman et al., 2025). Complementary white-box work probes XLT through the model's internal structure: cross-lingual neuron overlap asks whether languages share feature-selective neurons (Wang et al., 2024a), neuron activation overlap measures whether parallel inputs recruit similar circuits (Hu et al., 2025a), and sub-network similarity compares the language-specific computation used for transfer (Yun et al., 2023). These require model access and rely on the interpretability of internal activations, making them complementary to black-box, behavioral metrics such as ours.

**XLT as Unsupervised Domain Adaptation.** We are motivated by recent works modelling XLT as an unsupervised domain adaptation problem (Piratla et al., 2025), where source and target languages define domains with shared task semantics. Several prior works on domain adaptation highlight that target performance cannot be interpreted independently of source competence (Ben-David et al., 2010; Mansour et al., 2023; Miller et al., 2021). We take inspiration from these studies to propose a more principled metric to measure XLT.

## 3 Target and Source Accuracies are Coupled

For all the models and datasets described in Section 5, we plotted the source and target accuracy values in Figure 1. Also shown in each plot is the best linear fit in red and the line of perfect transfer, shown as a dashed line. The figure helps illuminate the nature of cross-lingual gaps at a glance, (1) target accuracy improves with source accuracy, and (2) target accuracy is often worse than that of source but not by much.

We observe that source and target accuracies follow a strong linear trend for all datasets and models (with the goodness of fit $R^2$ exceeding 0.94). This phenomenon is reminiscent of out-of-distribution accuracy falling on a line with in-distribution accuracy (Barbu et al., 2019; Miller et al., 2021; Salaudeen and Hardt, 2025). It is surprising that the strong linear trend continues to hold despite no clean separation of train-test splits and across various model architectures.

Since source and target accuracies are coupled, we cannot declare progress in XLT through source or target accuracy in isolation. An ideal measure of improvement in transfer must discount the improvements to target accuracy arising from general improvements in source accuracy. In the next section, we derive the standard measures from first principles, explain why they are misaligned for measuring XLT progress, and introduce our proposed metric.

## 4 Proposed XLT Metric: HAT Score

We denote by random variables $S, T_\ell \in [0, 1]$ the model's instance-level source and target language $\ell$ pass-rates, respectively. Pass-rate is the fraction of times the model responds correctly, for instance the pass-rate (@10) of 0.3 indicates the model is only correct three of ten times on a particular prompt. We sometimes refer example-level pass-rates also by accuracy. When a model achieves language parity, the model is expected to achieve comparable pass-rates in both source and target, i.e., $\mathbb{E}[T_\ell \mid S = s] = s, \forall \ell, s \in [0, 1]$.

We will formally define standard measures of XLT and argue why they do not indicate true progress in XLT.

**(1) Average Target Accuracy.** The average target accuracy for a given model $M$ on a dataset $D$ is expressed as $\mathbb{E}_{D,\ell}[T_\ell|M]$, which can also be written as a weighted aggregation of conditional expectations as shown below.

$$\begin{aligned}&\text{Avg. target accuracy} \triangleq \mathbb{E}_D[T \mid M]\\&= \mathbb{E}_S[\mathbb{E}_{D,\ell}[T_\ell \mid S, M]]\\&= \int_0^1 \mathbb{E}_{D,\ell}[T_\ell \mid S = s, M] p_S(s \mid M)\, ds \quad (1)\end{aligned}$$

For the sake of argument, consider two models: $M_1, M_2$ with comparable source conditional target accuracies, i.e., $\mathbb{E}[T \mid S = s, M_1] = \mathbb{E}[T \mid S = s, M_2], \quad \forall s \in [0, 1]$ but with the difference that $p_S(s \mid M_1)$ is shifted to the right of $p_S(s \mid M_2)$. Since the target accuracy is known to increase with source accuracy (Figure 1), the target accuracy for $M_1$ will be greater than that of $M_2$, despite comparable conditional accuracies.

**(2) XLT Gap.** Another standard measure of XLT progress is the difference between averaged source and target accuracies (XLT gap). This measure also suffers from the same issues as average target accuracy, as we show below.

$$\begin{aligned}&\text{XLT gap} \triangleq \mathbb{E}_{D,\ell}[S - T_\ell \mid M]\\&= \int_0^1 (s - \mathbb{E}_{D,\ell}[T_\ell \mid S = s, M]) p_S(s \mid M)\, ds.\end{aligned}$$

Following the same argument outlined for the average target accuracy, if we consider two models $M_1, M_2$ such that $s - \mathbb{E}[T \mid S = s, M_1] = s - \mathbb{E}[T \mid S = s, M_2]$ but with higher source accuracy for $M_1$, i.e., $p_S(s \mid M_1)$ is tilted right. The XLT gap, assuming the conditional gap $s - \mathbb{E}[T \mid S = s, M]$ is varies with $s$, will show variations between $M_1$ and $M_2$, despite their comparable conditional accuracies.

**(3) Transfer Score.** The transfer score was proposed by Goldman et al. (2025) to quantify the proportion of target questions answered correctly among questions that were answered correctly in the source language.

We express the score using our framework by deriving the fraction of examples correct in source and fraction correct in both source and target. For any given value of $S = s$ and target language $\ell$, the fraction of correct source responses is $s$, and the fraction of examples correct in both source and target is $s \times \mathbb{E}[T_\ell \mid S = s, M]$. Integrating over the full spectrum of $S$, the theoretical transfer score

($\tau_\ell$) is formulated as:

$$\tau_\ell \triangleq \frac{\int_0^1 p_S(s \mid M)\mathbb{E}[T_\ell \mid S = s, M]s\,ds}{\int_0^1 p_S(s \mid M)s\,ds}.$$

$$\tau = \mathbb{E}_\ell[\tau_\ell]$$

We will explain the (erroneous) sensitivity of $\tau_\ell$ with source accuracy with an example. Consider a model with perfect transfer, i.e., $\mathbb{E}[T_\ell \mid S = s, M] = s, \quad \forall s \in [0, 1]$. Depending on the distribution $p_S(s)$, the $\tau_\ell$ value can vary between 2/3 for uniformly distributed $p_S(s)$ and 1 when source accuracy is near perfect.

We used the above form of $\tau$ for drawing a conceptual insight. In practice, we estimate Transfer score as the average of transfer score per language, which is the ratio of number of examples correct in both source and target language to number of examples correct in source language.

We propose a new XLT metric called the **Hardness Adjusted Transfer Score (HAT Score)**, that captures XLT accurately and without confounders.

### 4.1 Transfer profile and HAT Score

**Transfer Profile.** A handy graphic is to plot the average target accuracy: $\mathbb{E}_{D,\ell}[T \mid S = s, M]$ on Y-axis and source pass-rate $s$ on X-axis. We refer to the regression curve defined by (s, $\mathbb{E}[T \mid S = s, M]$) as the transfer profile. If the model has perfect transfer then the transfer profile aligns with the y=x line. If the transfer profile for model A falls strictly below that of model B, we may claim better transfer for model B.

**HAT Score.** We propose to replace $p(s)$ in Equation 1 with a uniform distribution and instead define the metric as normalized expected target accuracy as below.

$$\text{HAT Score} = 2\int_0^1 \mathbb{E}_{D,\ell}[T_\ell \mid S = s, M]\,ds \quad (2)$$

The multiplier two ensures that the HAT score is bounded in the range [0, 1] and is comparable to other metrics.

**Empirical estimation of HAT score.** HAT score computation, unlike standard measures, requires an estimate for the conditional: $\mathbb{E}[T_\ell \mid S, M]$. For any given example $x^{(i)}$ from the dataset, we denote the instantiation of the same prompt in source and target languages as $x_s^{(i)}, x_\ell^{(i)}$ respectively. We approximate $S$ and $T_\ell$ for a given example with pass-rate across $N = 10$ promptings with model $M$ of their corresponding prompts $x_s^{(i)}, x_\ell^{(i)}$. We fit a linear regressor on all the estimated values of $(S, T_\ell)$ in order to approximate $\mathbb{E}[T_\ell \mid S, M]$. We use a linear regressor for fitting $T_\ell$ from S because (a) it works reasonably well, goodness of fit ($R^2$) was usually above 0.6 (some example plots are shown in Appendix A.5) (b) low-variance linear estimator can better handle sparse observations. We further impose a constraint on the estimator that when source language performance hits the minimum value of zero, i.e., $S = 0$, then $T$ must follow suite, i.e., $T_\ell = 0$. The estimation procedure is summarized below. For the linear predictor to pass-through (0, 0), the intercept must be zero.

$$((x_s^{(i)}, y_s^{(i)}), (x_\ell^{(i)}, y_\ell^{(i)})) \sim D$$

$$s^{(i)} = \sum_{n=1}^{N} \frac{\mathbb{J}(M(x_s^{(i)})_n, y_s^{(i)})}{N}$$

$$t^{(i)} = \mathbb{E}_\ell\left[\sum_{n=1}^{N} \frac{\mathbb{J}(M(x_\ell^{(i)})_n, y_\ell^{(i)})}{N}\right]$$

$$\hat{m} = \arg\min_m \sum_i \left(t^{(i)} - ms^{(i)}\right)^2 \quad (3)$$

$$\text{HAT Score} = 2\int_0^1 \hat{m}s\,ds = \hat{m} \quad (4)$$

where $M$ denotes the model being studied, $\mathbb{J}$ denotes the judge evaluator we use, and $M(x)_n$ denotes the model's response for the $n^{\text{th}}$ pass for input $x$. [1]

## 5 Experimental details

### 5.1 Models

We evaluate 20 models spanning five families: `Gemini` (Comanici et al., 2025; Google DeepMind, 2025a) (5), `Gemma` (Team et al., 2025; Google DeepMind, 2026b) (8), `Qwen` (2) (Yang et al., 2025), `GPT-OSS` (2) (OpenAI et al., 2025), and `Claude` (Anthropic, 2025) (3), to ensure our findings generalize across architectures, scales, and access paradigms (open-weight vs. proprietary API). The `Gemma-3` series (1B, 4B, 12B, 27B) provides a dense scaling ladder for studying how XLT evolves with model size within a single architecture. `Gemma-4` extends this with both dense (E2B, E4B, 31B) and mixture-of-experts (26B-A4B) variants. `Gemini-3-Flash` is evaluated at three reasoning depths (default, low, minimal) to isolate the effect of thinking budget on XLT. The `Claude` family

[1] For the linear fit, we used the standard numpy library in Python (Harris et al., 2020).

(Haiku 4.5, Sonnet 4.6, Opus 4.7) and GPT-OSS (20B, 120B) provide further independent scaling axes along model size and thinking budget. Our model selection also spans old-to-new timelines between 2024 to the latest 2026 Gemma-4 release.

### 5.2 Datasets

We evaluate on three multilingual benchmarks, each targeting a distinct capability: ECLeKTic (cross-lingual factoid QA) (Goldman et al., 2025), MGSMv2 (multilingual math reasoning), and MMLU-ProX-Lite (multilingual multiple-choice probing conceptual knowledge and require reasoning) (Xuan et al., 2025). The original MGSM benchmark (Shi et al., 2022) is known to contain translation and validation-related artifacts challenging accurate estimation of cross-lingual gaps. Peter et al. (2025) released a version fixing many issues with MGSM, which we use for our evaluation and refer to it as MGSMv2.

**Source and Target definition for MGSMv2, MMLU-ProX-Lite.** For both MGSMv2 and MMLU-ProX-Lite, we consider English as the source language and the target split is comprised of all available non-English languages in the dataset. Despite no clean train-test separation with LLMs, we believe English as source language is justified given the dominance of reasoning and/or conceptual knowledge focus of the two datasets. MGSMv2 contains ten non-English languages while MMLU-ProX-Lite contains twenty eight. Please see Appendix A.2 for the full list.

**Source and Target for ECLeKTic.** Source-target separation is more clearly specified in ECLeKTic. The dataset collects questions from Wikipedia pages that are only available in one language. For each factual question, the source is the language of the original Wikipedia article while target comprises of human-verified translations of the question and response into eleven other languages.

### 5.3 Experimental Setup

**Response Validation.** Multiple choice questions in MMLU-ProX-Lite can be validated with exact match. On the other hand, ECLeKTic and MGSMv2 require an LLM-based judge (Gu et al., 2025) for validation. We mark responses via majority voting with a model ensemble: Gemini-3-Flash (Low), GPT-OSS-20B, and Claude-Sonnet-4.6. For response validation, we prompt the judge model with the question, response and correct answer; this is a much simpler task than answering the original question. We employed an ensemble of judge models instead of one to suppresses for any self-preferential bias. We provide complete details of the evaluation protocol, including the answer extraction and answer scoring prompts in Appendix A.2.

**Inference Related Parameters.** When validating different models, we retain a common temperature (0.7) for all. We also retain default the thinking budget (unless explicitly specified) and sampling strategy. We provide dataset statistics and any further details in Appendix A.2.

### 5.4 Statistical Significance

All metrics are accompanied by 95% confidence intervals estimated using bootstrap method (Efron, 1979). We resample with replacement ($1,000$ times) from the observations to estimate the distribution of the aggregated measure, which yields their mean and confidence interval.

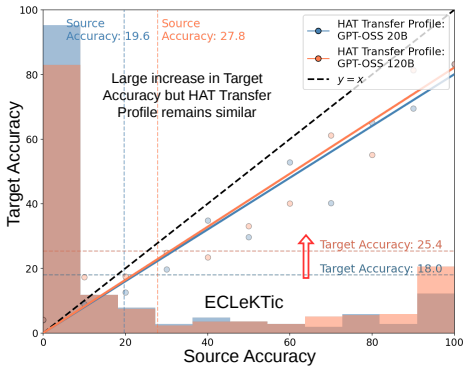
## 6 Results

We present comprehensive cross-lingual evaluation results in Tables 2 (ECLeKTic), 3 (MMLU-ProX-Lite) and 4 (MGSMv2) of Section A.1 in Appendix. We establish the reliability of HAT score in Section 6.1. We then utilize HAT score to address interesting questions related to progress. We track the improvements on XLT with model size in Section 6.2, over time in 6.3, and conclude with an overall comparison across all models in 6.4.

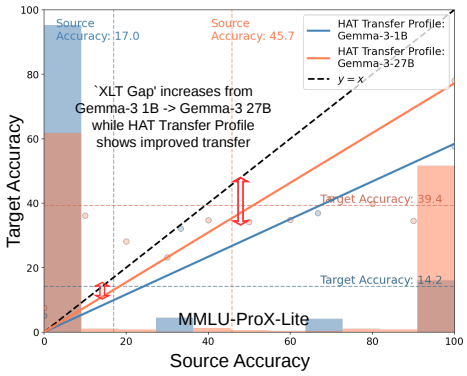
### 6.1 HAT Score Reliability

In Figure 2, we present two comparisons between different pairs of models. In each of the plots, we show transfer profile: $\mathbb{E}[T \mid S]$ for both the models. At the bottom of each plot, we show frequency histogram of source accuracy.

**HAT Score is Preferable over Target Accuracy.** In Figure 2 (a), we contrast GPT-OSS 20B with GPT-OSS 120B on ECLeKTic. Target accuracy for the 120B model (25.4%) is significantly larger than that of 20B model (18.0%) despite overlapping transfer profiles, i.e., no improvement in XLT. The increase in target accuracy with 120B is due to a right shifted population as visualized by the frequency histogram. HAT score for the two models is close to 0.82, thus correctly indicating no improvement in transfer.

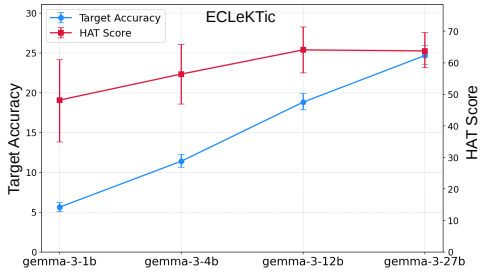



(a) Target accuracy indicates progress despite overlapping transfer profiles. HAT score remains unchanged, thereby better representing progress.

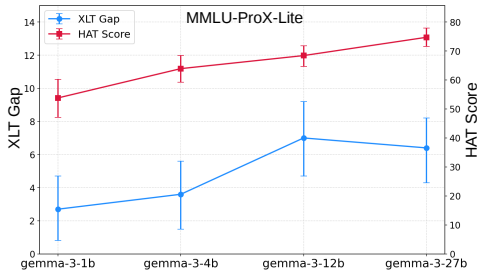



(b) XLT gap falsely indicates worsening transfer although there is a clear improvement on transfer profile.

Figure 2: **Improvements in source accuracy inflate traditional metrics like average target accuracy (a) and XLT gap (b).** Please refer Section 6.1 for more details.




(a) ECLeKTic: Target accuracy inflates while true transfer (HAT) plateaus.




(b) MMLU-ProX-Lite: XLT gap falsely implies regression as models scale.

Figure 3: **Scaling ladder with HAT score.** Traditional metrics distort XLT progress with model size: (a) target accuracy falsely inflates on ECLeKTic, and (b) the XLT gap falsely shows regression on MMLU-ProX-Lite. The HAT score corrects the confounders, revealing that actual transfer progress is much more gradual (Section 6.2).

**HAT Score is Preferable over Transfer Score** In the same Figure 2a, transfer score unlike HAT score showed a significant jump from 53.5 to 61 between GPT-OSS models despite overlapping transfer profiles. We contrast transfer and HAT Scores for various models on ECLeKTic in Table 14 of Appendix A.6, where we discuss more differences in detail.

**HAT Score is Preferable over XLT gap.** Figure 2 (b) contrasts the `Gemma-3-1B` model with the `Gemma-3-27B` model on MMLU-ProX-Lite. The 27B model clearly improves over the 1B model, which is apparent from its steeper slope. However, the XLT gap increases from 2.7% for the 1B model to 6.4% for the 27B model, falsely indicating worsened XLT. The HAT score, which represents the area under the transfer profile, correctly captures this progress by increasing from 58.4 to 77.2.

In all the examples above, HAT score is robust to the shifting frequency histogram of source accuracies. Overall, HAT score is a more reliable indicator of XLT progress.

### 6.2 XLT Progress with Model Size Scaling

We compare progress with model size among models that belong to the same family to control for any unrelated changes to training recipes. Figure 3 compares across the `Gemma-3` class of models. Progress indicated by HAT score in contrast to target accuracy is much slower. This seemingly slow progress is in part because the cross-lingual transfer in the smallest model is not as bad as indicated by target accuracy. Nevertheless, we continue to see some progress with model size. The observation aligns with the general trend of improved generalization with larger models (Zhang et al., 2016).

**XLT in Small Models.** Our validation includes some relatively small size models in the range of 1B to 4B size: `Gemma-3-1B`, `Gemma-3-4B`, `Qwen-3-4B`, `Gemma-4-E2B`. We observe from Figures 3, 5 that language transfer is not broken at very

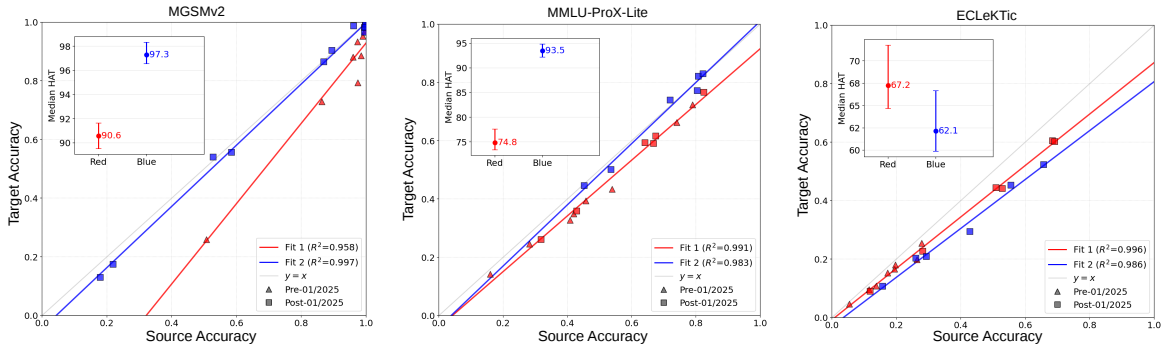



Figure 4: We show pair of line fit for accuracy observations of Figure 1. The blue line corresponds to newer models with a knowledge cutoff date after January 2025. For both MGSMv2, MMLU-ProX-Lite, newer models shift to a separate, tighter linear regime. Moreover for the two benchmarks, the median HAT score of recent models exceeds that of older models, showing a clear progress of cross-lingual reasoning transfer over time (Section 6.3).

small scales. The HAT score for the four models on the three datasets is in 50-80 range meaning the target accuracy is 50-80% of the source accuracy. Target accuracy, on the other hand, as shown in Figure 3 provides an overly pessimistic estimate for small models.

### 6.3 XLT Progress over Time

Assuming we made progress with time, we expect significant deviations to our accuracy-on-the-line observation of Figure 1. All the new models, if they are remarkably better than the old, must fall closer to the perfect transfer line while deviating from the transfer profile of older models. If that were the case, we expect to see a pair of lines fitting the accuracy observations better than a single line, and a clear separation between old and new models falling on two lines.

Our analysis is inspired from Dominguez-Olmedo et al. (2025). We sort models by their knowledge cutoff date and group models into old and new for various possible dates. Moreover, progress need not percolate uniformly across new models, so we must allow a subset of new models falling in line with old models. We employ a latent binary variable indicating if a new model must be placed in line with old or new models. For a given cut-off date, we optimize the latent variable assignment for all new models with a combinatorial subset search algorithm to minimize the overall mean-squared error. Finally, we fit two lines for (a) selected new models, (b) unselected new model and old models. Please refer to Appendix A.3 for more details. We consider as progress if the accuracy line for (a subset of) newer models is closer to perfect transfer and their median HAT score is significantly better than that of old models.

The pair-of-lines fit for all the datasets is shown in Figure 4. We observe a significant improvement with time for both MGSMv2 and MMLU-ProX-Lite. The median hat score for MGSMv2 (0.90→0.97) and MMLU-ProX-Lite (0.75→0.93) show significant improvement with time. But we do not see such a trend with ECLeKTic. The near-perfect HAT score for new models for MGSMv2, MMLU-ProX-Lite indicate they are nearly solved.

### 6.4 Overall Comparison

The full chart of HAT scores across all models and all datasets shown in Figure 5 reveal Gemini-3-Flash (with default thinking level) as the clear winner of cross-lingual transfer. Despite its multilingual emphasis, neither Gemma-3 or Gemma-4 fared better.

Between Gemma models of comparable size: `Gemma-3-27B`, `Gemma-4-26B-A4B`, HAT score over all datasets is comparable. However, between models of comparable size but different model families: `GPT-OSS-20B`, `Gemma-4-31B`, `Qwen-30B-A3B`, we do not see comparable differences in HAT scores. This trend across model families indicates that pretraining recipes are likely more important for HAT score than model size.

### 6.5 Transfer Between Shared and Different Scripts

Is transfer across languages with a shared script better? Recall that HAT score quantifies transfer between source and any subset of target languages. We split the target languages into two non-overlapping subsets based on whether they share the script with source. Transfer between languages

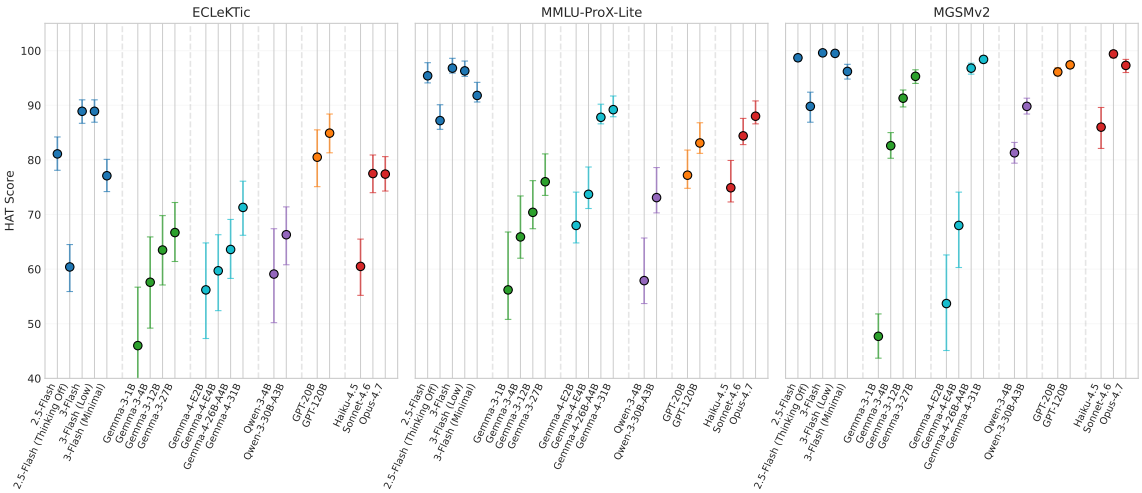



Figure 5: HAT scores for all models and all datasets. We observe transfer improvements with model size, thinking and model upgrades (Gemma3→Gemma4, Gemini-2.5→Gemini-3). Please see Section 6.4 for detailed description.

with script differences is commonly understood to be weaker than that of transfer between shared script languages. Intuitively, using the same alphabet (like Latin) helps align words, while switching scripts (like Latin to Arabic or Cyrillic) adds a barrier. Table 1 breaks down the average HAT scores by script similarity.

| Dataset | HAT Score Same Script | HAT Score Different Script | Δ (Delta) |
|---|---|---|---|
| ECLeKTic | $79.1 \pm 10.7$ | $61.7 \pm 14.9$ | $\mathbf{+17.4 \pm 7.4}$ |
| MGSMv2 | $89.6 \pm 14.0$ | $89.1 \pm 15.3$ | $+0.5 \pm 2.9$ |
| MMLU-ProX-Lite | $79.0 \pm 12.0$ | $80.9 \pm 12.4$ | $-1.9 \pm 1.4$ |

Table 1: XLT Progress based on Script Congruence between Source and Target. We show aggregated HAT Score across all models. Cross-script transfer is weak only in ECLeKTic.

We observe a clear depreciation in HAT score from same to different script only on ECLeKTic. We do not see any trend for the other two tasks. ECLeKTic requires recalling factual knowledge and reconciling entities in any language to that of source language. The other two tasks, on the other hand, only require conceptual reasoning.

## 7 Discussion and Future Work

In this work, we presented a reliable measure of cross-lingual transfer strength and used it to quantify the influence of various modeling choices on transfer. We further demonstrated that modern LLMs have made remarkable progress on the two reasoning-like datasets: MGSMv2, MMLU-ProX-Lite, while progress on ECLeKTic is stalled. We hope that our metric and insights tease out factors that positively influence transfer and further catalyze research on cross-lingual transfer.

**Role of thinking.** Comparison between Gemini-2.5-Flash and Gemini-3-Flash model with and without thinking in Figure 5 indicate that thinking helps improve XLT on all benchmarks including ECLeKTic. When the prompts are short and source is English, we may intuitively reason that the model can translate the prompts to English and then respond, effectively bridging the cross-lingual gap. We did not observe any explicit behavior of translating the whole prompt to English and then responding. We see longer thinking in non-English indicating that some number of thinking tokens are allocated for language bridging. Empirical observations with Gemini models suggest 5% additional thinking length for prompts in target language on MGSMv2 and about 15-20% on other datasets (Appendix A.8). Transfer scores must ideally also quantify parity on (test-)compute normalized accuracies between source and target languages because financial cost also influences technology access.

**Call for Action.** MGSMv2 and MMLU-ProX-Lite are popular and are nearly solved after January 2025 (Section 6.3). The underlying tasks for each of them: GSM8K (Cobbe et al., 2021)/MMLU-Pro (Wang et al., 2024b), are relatively old. Search-off factual recall such as ECLeKTic is of less practical significance because search is easily available. Besides, the usage pattern of models have evolved from these standard datasets but the progress on multilingual benchmarks has been slow. This calls for more challenging practically motivated benchmarks that go beyond single-step evaluations where we are likely to see large gaps due to compounding cross-lingual errors from multiple sub-problems.

## 8 Limitations.

**Data Contamination.** HAT Score like any other measure is not robust to data contamination and dataset errors. New models tracing closely the line of perfect transfer in Figure 1 could also indicate contamination without algorithmic enhancements. Thereby, we risk conflating algorithmic progress with contamination (due to leakage and memorization) like with any other metric.

**Measurement Error.** The empirical estimation of accuracy uses a finite sample of promptings (e.g., $N = 10$) and this inherently introduces some sampling variance. Our evaluation framework leverages 1000 bootstrap iterations alongside a sufficiently high number of inference passes to effectively alleviate this measurement error.

## References


Anthropic. 2025. Claude family. Accessed: May 21, 2026.

Anthropic. n.d. Anthropic's Transparency Hub. https://www.anthropic.com/transparency. Accessed: 2026-05-25.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 4623–4637. Association for Computational Linguistics.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *Preprint*, arXiv:2007.07834.

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

Cohere Labs. 2024. The ai language gap.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. 2025. Training on the test task confounds evaluation and emergence. *Preprint*, arXiv:2407.07890.

B. Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *Preprint*, arXiv:2502.21228.

Google. 2026a. Gemini 2.5 Flash. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash. Google AI for Developers documentation. Knowledge cutoff: January 2025.

Google. 2026b. Gemini 3 Developer Guide. https://ai.google.dev/gemini-api/docs/gemini-3. Google AI for Developers documentation. Knowledge cutoff: January 2025.

Google DeepMind. 2025a. Gemini 3 [large language model]. Official Blog Release.

Google DeepMind. 2025b. Gemma 3 model card. https://ai.google.dev/gemma/docs/core/model_card_3. Google AI for Developers model card.

Google DeepMind. 2026a. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Google AI for Developers model card. Training data cutoff: January 2025.

Google DeepMind. 2026b. Gemma 4: Open models family.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. 2026. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *Preprint*, arXiv:2509.07968.

Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. 2025. Mubench: Assessment of multilingual capabilities of large language models across 61 languages. *Preprint*, arXiv:2506.19468.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, and 7 others. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

Amey Hengle, Prasoon Bajpai, Soham Dan, and Tanmoy Chakraborty. 2024. Multilingual needle in a haystack: Investigating long-context behavior of multilingual large language models. *Preprint*, arXiv:2408.10151.

Finn Hillebrandt. 2026. Knowledge cutoff date: Definition & explanation. `https://www.gradually.ai/en/ai-glossary/knowledge-cutoff-date/`. Gradually.ai.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Preprint*, arXiv:2003.11080.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025a. Large language models are cross-lingual knowledge-free reasoners. *Preprint*, arXiv:2406.16655.

Songbo Hu, Ivan Vulić, and Anna Korhonen. 2025b. Quantifying language disparities in multilingual large language models. *Preprint*, arXiv:2508.17162.

Maxim Ifergan, Leshem Choshen, Roee Aharoni, Idan Szpektor, and Omri Abend. 2024. Beneath the surface of consistency: Exploring cross-lingual knowledge representation sharing in llms. *Preprint*, arXiv:2408.10646.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2023. Domain adaptation: Learning bounds and algorithms. *Preprint*, arXiv:0902.3430.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. *Preprint*, arXiv:2107.04649.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 5 others. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Jan-Thorsten Peter, David Vilar, Tobias Domhan, Dan Malkin, and Markus Freitag. 2025. Mind the gap... or not? how translation errors and evaluation details skew multilingual results. *Preprint*, arXiv:2511.05162.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Vihari Piratla, Purvam Jain, Darshan Singh, Partha Talukdar, and Trevor Cohn. 2025. Rethinking cross-lingual gaps from a statistical viewpoint. *Preprint*, arXiv:2510.15551.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Olawale Salaudeen and Moritz Hardt. 2025. Imagenot: A contrast with imagenet preserves model rankings. *Preprint*, arXiv:2404.02112.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

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, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Hetong Wang, Pasquale Minervini, and Edoardo M. Ponti. 2024a. Probing the emergence of cross-lingual alignment during llm training. *Preprint*, arXiv:2406.13229.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11).

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, and 13 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

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

Taejun Yun, Jinhyeon Kim, Deokyeong Kang, Seong Hoon Lim, Jihoon Kim, and Taeuk Kim. 2023. X-sns: Cross-lingual transfer prediction through subnetwork similarity. *Preprint*, arXiv:2310.17166.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *Preprint*, arXiv:2402.18815.

| Model | Target Accuracy (%) (↑) | | XLT Gap (%) (↓) | | HAT Score (↑) | |
|---|---|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI |
| GEMINI-2.5-FLASH | 44.5 | [41.1, 48.1] | 6.4 | [4.3, 8.5] | 81.7 | [78.6, 84.7] |
| GEMINI-2.5-FLASH THINKING-OFF | 29.4 | [26.5, 32.5] | 13.3 | [10.5, 16.3] | 60.4 | [55.9, 64.8] |
| GEMINI-3-FLASH | 60.5 | [56.8, 64.3] | 8.0 | [5.6, 10.4] | 84.6 | [82.1, 87.1] |
| GEMINI-3-FLASH LOW | 60.2 | [56.5, 64.0] | 8.9 | [6.5, 11.2] | 84.0 | [81.6, 86.7] |
| GEMINI-3-FLASH MINIMAL | 52.3 | [48.9, 56.1] | 13.4 | [10.2, 16.3] | 74.2 | [70.9, 77.5] |
| GEMMA-3-1B | 4.6 | [3.5, 6.0] | 0.8 | [-0.3, 2.2] | 48.2 | [34.9, 61.0] |
| GEMMA-3-4B | 9.6 | [7.7, 11.8] | 1.9 | [0.3, 4.0] | 56.4 | [47.0, 65.9] |
| GEMMA-3-12B | 15.2 | [12.8, 17.9] | 2.1 | [-0.2, 4.3] | 64.1 | [56.8, 71.4] |
| GEMMA-3-27B | 19.8 | [17.2, 23.0] | 6.5 | [4.1, 8.9] | 63.8 | [58.4, 69.6] |
| QWEN-3-4B | 10.8 | [8.9, 13.1] | 2.9 | [1.1, 5.1] | 56.5 | [47.8, 64.9] |
| QWEN-3-30B-A3B | 16.6 | [14.1, 19.5] | 2.8 | [0.7, 5.1] | 65.9 | [59.7, 72.6] |
| GPT-OSS-20B | 18.0 | [15.3, 21.2] | 1.6 | [0.2, 3.2] | 80.1 | [74.7, 85.3] |
| GPT-OSS-120B | 25.4 | [22.4, 28.7] | 2.4 | [0.6, 4.1] | 82.2 | [77.9, 86.1] |
| GEMMA-4-E2B-IT | 9.0 | [7.2, 11.0] | 2.8 | [0.9, 4.9] | 51.8 | [42.4, 61.3] |
| GEMMA-4-E4B-IT | 10.7 | [8.7, 13.0] | 5.0 | [3.1, 6.8] | 58.4 | [51.2, 65.7] |
| GEMMA-4-26B-A4B-IT | 20.3 | [17.4, 23.3] | 5.7 | [3.0, 8.3] | 63.9 | [57.9, 69.6] |
| GEMMA-4-31B-IT | 22.6 | [19.5, 26.0] | 5.5 | [3.1, 8.0] | 68.6 | [62.9, 74.1] |
| CLAUDE-HAIKU-4.5 | 20.8 | [17.8, 23.9] | 8.4 | [5.9, 11.3] | 59.8 | [53.8, 65.7] |
| CLAUDE-SONNET-4.6 | 44.1 | [40.2, 48.0] | 8.7 | [5.6, 11.5] | 74.7 | [70.9, 78.5] |
| CLAUDE-OPUS-4.7 | 45.2 | [41.6, 49.1] | 10.3 | [7.4, 13.5] | 73.5 | [69.7, 77.2] |

Table 2: **Comprehensive evaluation results for ECLeKTic.** Complete estimates and 95% confidence intervals for Average Target Accuracy, XLT Gap, and HAT score across all tested model configurations.

# A Appendix

## A.1 Evaluations

In this section, we present the comprehensive evaluation results across all models and model families investigated in our study. Tables 2, 3, and 4 detail the complete metric information—including the expected Target Accuracy, the Cross-Lingual Transfer (XLT) Gap, and our proposed HAT Score, along with their respective 95% confidence intervals for the ECLeKTic, MMLU-ProX-Lite, and MGSMv2 datasets.

| Model | Target Accuracy (%) (↑) | | XLT Gap (%) (↓) | | HAT Score (↑) | |
|---|---|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI |
| GEMINI-2.5-FLASH | 73.4 | [73.3, 75.5] | -1.9 | [-3.4, -0.5] | 97.5 | [96.2, 98.8] |
| GEMINI-2.5-FLASH THINKING-OFF | 61.6 | [60.1, 62.2] | 5.9 | [4.5, 7.4] | 88.9 | [87.1, 90.3] |
| GEMINI-3-FLASH | 82.1 | [81.6, 82.6] | -1.4 | [-2.6, -0.3] | 98.8 | [97.8, 99.9] |
| GEMINI-3-FLASH LOW | 83.0 | [82.5, 83.5] | -0.8 | [-2.0, 0.4] | 98.3 | [97.3, 99.3] |
| GEMINI-3-FLASH MINIMAL | 77.2 | [76.6, 77.8] | 3.2 | [1.9, 4.6] | 93.5 | [92.1, 94.8] |
| GEMMA-3-1B | 14.2 | [13.7, 14.7] | 2.7 | [0.8, 4.7] | 58.4 | [52.3, 64.1] |
| GEMMA-3-4B | 24.6 | [23.9, 25.2] | 3.6 | [1.5, 5.6] | 67.5 | [63.1, 71.7] |
| GEMMA-3-12B | 34.9 | [34.3, 35.6] | 7.0 | [4.7, 9.2] | 71.6 | [68.2, 74.7] |
| GEMMA-3-27B | 39.4 | [38.6, 40.1] | 6.4 | [4.3, 8.2] | 77.2 | [74.4, 80.1] |
| QWEN-3-4B | 32.7 | [32.0, 33.4] | 8.1 | [6.0, 10.3] | 69.3 | [65.8, 72.6] |
| QWEN-3-30B-A3B | 43.4 | [42.7, 44.2] | 10.5 | [8.5, 12.4] | 74.8 | [71.9, 77.6] |
| GPT-OSS-20B | 66.3 | [65.7, 66.9] | 7.7 | [6.5, 9.0] | 88.9 | [87.6, 90.3] |
| GPT-OSS-120B | 72.3 | [71.8, 73.0] | 6.6 | [5.4, 8.0] | 90.6 | [89.2, 92.0] |
| GEMMA-4-E2B-IT | 26.1 | [25.5, 26.7] | 5.7 | [3.2, 8.6] | 59.3 | [54.7, 63.5] |
| GEMMA-4-E4B-IT | 35.8 | [35.2, 36.5] | 7.1 | [5.1, 9.0] | 74.8 | [71.6, 77.8] |
| GEMMA-4-26B-A4B-IT | 48.7 | [48.0, 49.5] | 4.8 | [2.6, 7.2] | 78.7 | [76.2, 81.1] |
| GEMMA-4-31B-IT | 59.4 | [58.6, 60.1] | 4.8 | [2.9, 6.8] | 84.7 | [82.5, 86.7] |
| CLAUDE-HAIKU-4.5 | 44.6 | [43.9, 45.3] | 0.6 | [-1.9, 3.1] | 76.5 | [73.5, 79.2] |
| CLAUDE-SONNET-4.6 | 59.5 | [58.7, 60.2] | 7.4 | [5.7, 9.0] | 85.4 | [83.6, 87.3] |
| CLAUDE-OPUS-4.7 | 76.7 | [76.2, 77.5] | 5.8 | [3.8, 7.3] | 89.3 | [87.8, 90.9] |

Table 3: **Comprehensive evaluation results for MMLU-ProX-Lite.** Complete estimates and 95% confidence intervals for Average Target Accuracy, XLT Gap, and HAT score across all tested model configurations.

| Model | Target Accuracy (%) (↑) | | XLT Gap (%) (↓) | | HAT Score (↑) | |
|---|---|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI |
| GEMINI-2.5-FLASH | 98.2 | [97.7, 98.6] | 1.3 | [0.8, 1.9] | 98.7 | [98.1, 99.2] |
| GEMINI-2.5-FLASH THINKING-OFF | 55.6 | [53.9, 57.5] | 2.8 | [-0.1, 5.4] | 89.8 | [86.9, 92.4] |
| GEMINI-3-FLASH | 99.3 | [99.1, 99.6] | 0.4 | [0.1, 0.8] | 99.6 | [99.3, 99.9] |
| GEMINI-3-FLASH LOW | 99.4 | [99.2, 99.6] | 0.5 | [0.2, 1.0] | 99.5 | [99.1, 99.8] |
| GEMINI-3-FLASH MINIMAL | 86.5 | [85.4, 87.7] | 0.4 | [-1.8, 2.5] | 96.2 | [94.8, 97.5] |
| GEMMA-3-1B | 25.9 | [24.6, 27.4] | 24.7 | [21.3, 28.3] | 47.7 | [43.7, 51.8] |
| GEMMA-3-4B | 72.9 | [71.5, 74.3] | 13.4 | [10.8, 15.9] | 82.6 | [80.3, 85.0] |
| GEMMA-3-12B | 88.1 | [87.1, 89.1] | 7.9 | [6.3, 9.7] | 91.3 | [89.7, 92.8] |
| GEMMA-3-27B | 93.3 | [92.6, 94.1] | 4.1 | [2.7, 5.4] | 95.3 | [94.0, 96.5] |
| QWEN-3-4B | 79.4 | [77.9, 80.8] | 18.1 | [16.2, 20.1] | 81.3 | [79.4, 83.2] |
| QWEN-3-30B-A3B | 88.6 | [87.6, 89.6] | 9.8 | [8.2, 11.4] | 89.8 | [88.4, 91.3] |
| GPT-OSS-20B | 95.2 | [94.6, 95.8] | 3.8 | [3.0, 4.6] | 96.1 | [95.4, 96.9] |
| GPT-OSS-120B | 97.0 | [96.4, 97.5] | 2.6 | [2.0, 3.2] | 97.4 | [96.8, 98.0] |
| GEMMA-4-E2B-IT | 13.0 | [11.9, 14.1] | 5.0 | [2.2, 8.0] | 53.7 | [45.1, 62.6] |
| GEMMA-4-E4B-IT | 17.5 | [16.3, 18.8] | 4.4 | [2.3, 6.8] | 68.0 | [60.3, 74.1] |
| GEMMA-4-26B-A4B-IT | 96.5 | [95.9, 97.1] | 3.0 | [1.9, 4.2] | 96.8 | [95.7, 97.7] |
| GEMMA-4-31B-IT | 98.3 | [97.8, 98.6] | 1.4 | [0.6, 2.1] | 98.4 | [97.8, 99.0] |
| CLAUDE-HAIKU-4.5 | 54.0 | [52.3, 55.9] | -1.3 | [-4.4, 2.2] | 86.0 | [82.1, 89.6] |
| CLAUDE-SONNET-4.6 | 98.7 | [98.3, 99.1] | -2.6 | [-4.9, -0.5] | 99.4 | [98.7, 100.0] |
| CLAUDE-OPUS-4.7 | 90.4 | [89.4, 91.5] | -1.0 | [-2.8, 1.1] | 97.3 | [96.0, 98.4] |

Table 4: **Comprehensive evaluation results for MGSMv2.** Complete estimates and 95% confidence intervals for Average Target Accuracy, XLT Gap, and HAT score across all tested model configurations.

### A.1.1 ECLeKTic HAT Plots

Detailed evaluation curves and HAT score transfer profiles for the ECLeKTic dataset are provided below.

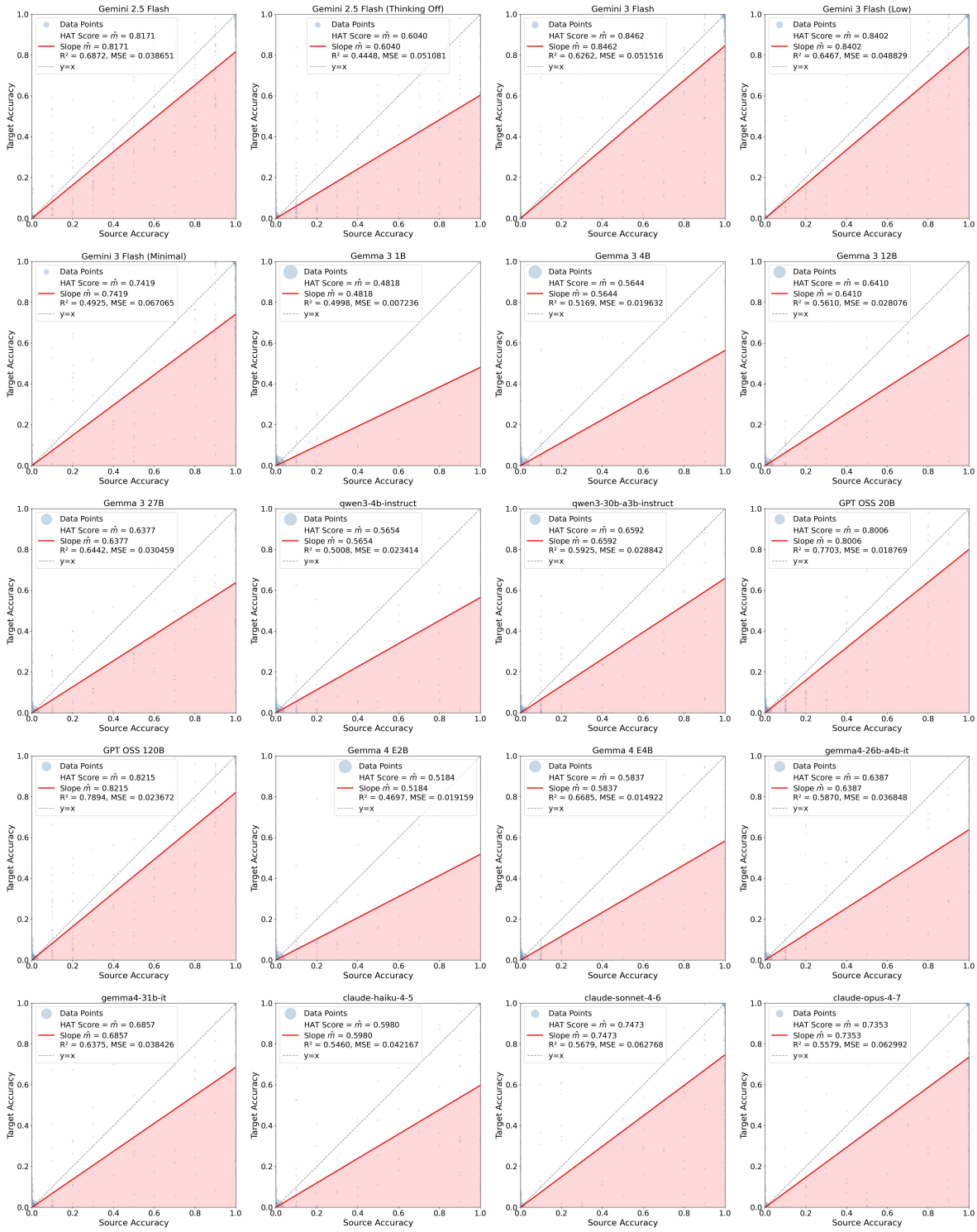



Table 5: HAT Profiles for ECLeKTic

### A.1.2 MGSMv2 HAT Plots

Detailed evaluation curves and HAT score transfer profiles for the MGSMv2 dataset are provided below.

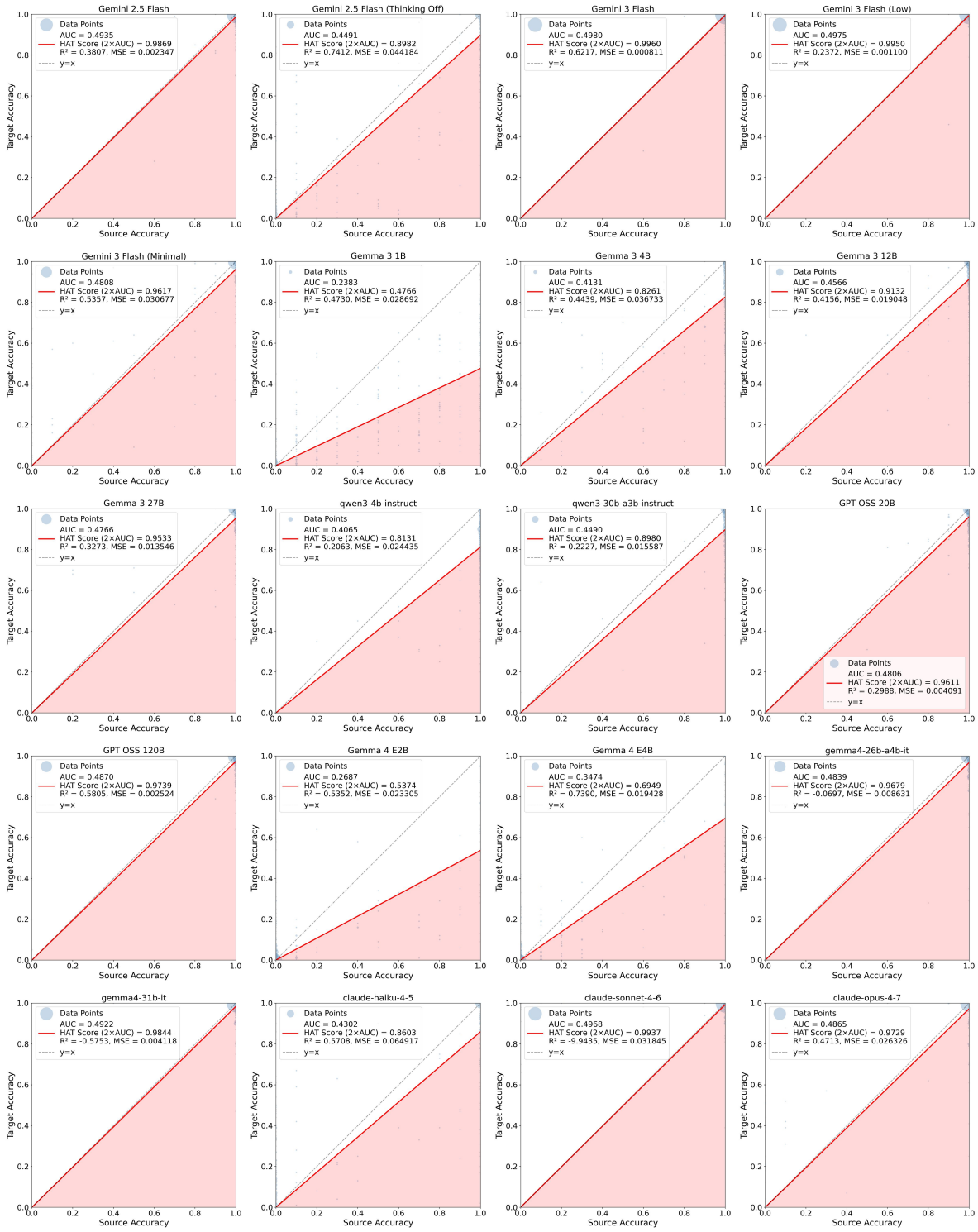



Table 6: HAT Profiles for MGSMv2

### A.1.3 MMLU-ProX-Lite HAT Plots

Detailed evaluation curves and HAT score transfer profiles for the MMLU-ProX-Lite dataset are provided below.

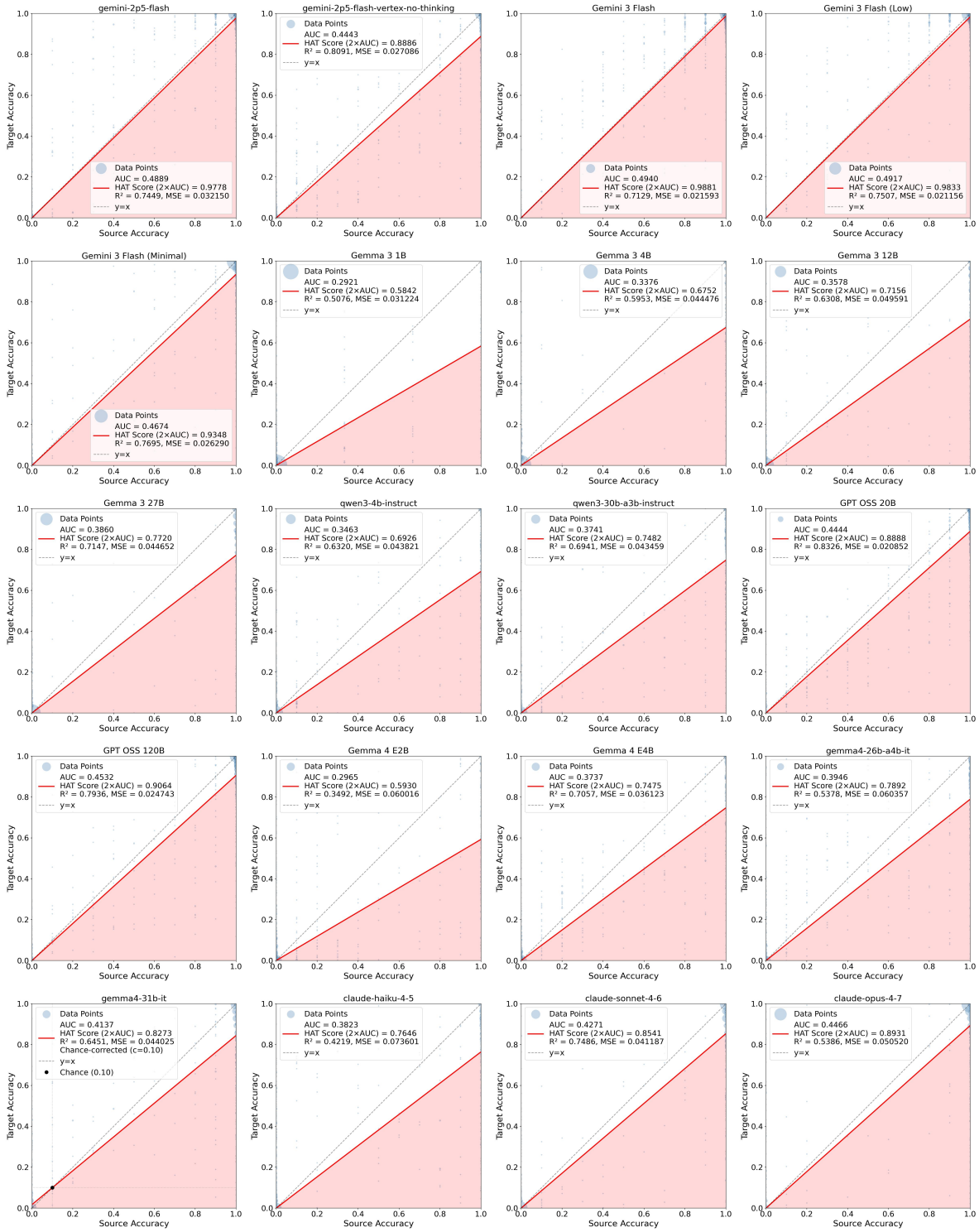



Table 7: HAT profiles for MMLU-ProX-Lite.

| Dataset | Languages |
|---|---|
| **MGSMv2** | English, Bengali, Chinese, French, German, Japanese, Russian, Spanish, Swahili, Telugu, Thai |
| **ECLeKTic** | English, Chinese, French, German, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Portuguese, Spanish |
| **MMLU-ProX-Lite** | English, Afrikaans, Arabic, Bengali, Chinese, Czech, French, German, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Marathi, Nepali, Portuguese, Russian, Serbian, Spanish, Swahili, Telugu, Thai, Ukrainian, Urdu, Vietnamese, Wolof, Yoruba, Zulu |

Table 8: Distribution of languages across datasets, highlighting the broader linguistic context of our evaluations.

## A.2 Experimental Details

We provide details for all datasets, models covered in our study. We also provide details of the answer extraction and evaluation protocols. [2]

### A.2.1 Datasets

We utilize a representative collection of datasets to study XLT from the perspective of both 'knowledge transfer' and 'reasoning transfer'. We present the exhaustive collection of languages represented in our study in Table 8. Below are the details of the datasets we use:[3]

- **ECLeKTic:** (Goldman et al., 2025) ECLeKTic is designed specifically for cross-lingual knowledge transfer in LLMs through closed-book QA. We use it to test whether knowledge available in a source language can be recovered in a target language, using an efficient transfer mechanisms across languages.

- **MGSMv2:** (Peter et al., 2025; Shi et al., 2022) MGSM evaluates multilingual mathematical reasoning by manually translating 250 GSM8K (Cobbe et al., 2021) grade-school math problems into 10 typologically diverse languages. It is useful for XLT because it tests whether reasoning skills transfer beyond English, including to underrepresented languages such as Bengali, Swahili, and Telugu. Recent works have reported (and corrected) translation errors in MGSM (Peter et al., 2025). To avoid confounding our observations with such artifacts, we use the corrected version proposed in the same study. We call this version MGSMv2.

- **MMLU-ProX-Lite:** (Xuan et al., 2025) MMLU-ProX extends MMLU-Pro to 29 languages. We identify as a midway scale for measuring XLT, between ECLeKTic (pure knowledge recall) and MGSMv2 (pure reasoning capacity). Furthermore, it supports up to 29 languages providing more support for experimental artifacts. We use the 'MMLU-ProX-Lite' split, having 588 questions per language, to facilitate cost-efficient evaluations.

### A.2.2 Models

We evaluate our approach using a diverse suite of both open-source and closed-source frontier large language models. All models used in our experiments were accessed via Vertex AI on Google Cloud Pantheon. The complete list of evaluated models is as follows:

- **Gemini Family:** Gemini-2.5-Flash, Gemini-2.5-Flash Thinking-Off, Gemini-3-Flash, Gemini-3-Flash Low, Gemini-3-Flash Minimal
- **Gemma Family:** Gemma-3-1B, Gemma-3-4B, Gemma-3-12B, Gemma-3-27B, Gemma-4-E2B-IT, Gemma-4-E4B-IT, Gemma-4-26B-A4B-IT, Gemma-4-31B-IT
- **Claude Family:** Claude-Haiku-4.5, Claude-Sonnet-4.6, Claude-Opus-4.7
- **Qwen Family:** Qwen-3-4B, Qwen-3-30B-A3B
- **GPT-OSS Family:** GPT-OSS-20B, GPT-OSS-120B

### A.2.3 Evaluation Protocol

The exact prompts used for extracting the answers from different models are provided in Figure 6.

[2] We acknowledge the use of AI tools for providing design-wise creative inspiration for some of the visualizations in this work, including Figure 4 and Figure 5.

[3] All datasets utilized in our study are publicly available under open-source licenses that permit academic research and evaluation.

```
ECLeKTic -
"""
You are a helpful assistant. Please answer the following question. Provide only the answer. Do not include any explanations or additional text.
Question: {question}
Answer:
"""

MGSMv2 -
"""
You are a helpful assistant. Please answer the following question."
Question: {question}
Answer:
"""


MMLU-ProX-Lite -
"""
Answer the following multiple choice question.
Reply with ONLY the letter of the correct option "
(e.g. A, B, C, ...).
Do not include any other text.
Question:{question}
Options:{options_str} # String of all options
Answer:"
"""
```

Figure 6: Answer extraction prompts used for different datasets: ECLeKTic (Top), MGSMv2 (Middle), MMLU-ProX-Lite (Bottom)

For evaluation, we employ methods specific to the response format and answer semantics of each dataset. The three datasets span distinct evaluation paradigms: LLM-as-judge (Gu et al., 2025) for open-ended factual QA (ECLeKTic) and math reasoning (MGSMv2), and deterministic letter extraction for multiple-choice QA (MMLU-ProX-Lite). For our LLM-as-judge based evaluations, we use k=3 models from different model families: `Gemini-3-Flash` (Low Thinking), `GPT-OSS 20B` and `Claude-Sonnet-4.6` as our judges. Each example is classified as correct or incorrect based on a majority voting of scores from the three judges. We provide the exact prompt used for judging the correctness of model predictions in Section A.2.4.

### A.2.4 Evaluation Judge Prompts

Figure 7 shows the judge prompt we use for ECLeKTic. We use the same prompt for factuality as is used in SimpleQA (Haas et al., 2026). For MGSMv2, we use a simple two-step based judgment prompt, that first extracts the model answer from the model prediction (that includes rationale) and then matches it against the ground truth. We provide this prompt in Figure 8.

### A.2.5 Judge Reliability

We create a split of ECLeKTic (denote as Year-ECLeKTic). This contains questions that are answered in years of format YYYY (eg. Question: When was Kreeda Bharati established? Answer: 1992). We evaluate all responses from all 20 models used in our study with both Exact-Match scorer and our LLM-as-a-Judge setup that uses majority score from `Gemini-3-Flash`, `Claude-Sonnet-4.6` and `GPT-OSS 20B`. Table 9 shows average accuracy obtained from two scoring methods across models. Moreover, >99.9% agreement is between scores provided by the two methods, showcasing high reliability of our evaluation protocol.

```
Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"].
First, I will give examples of each grade, and then you will grade a new example.


The following are examples of CORRECT predicted answers.
‘‘‘
Question: What are the names of Barack Obama’s children?
Gold target: Malia Obama and Sasha Obama
Predicted answer 1: sasha and malia obama
Predicted answer 2: most people would say Malia and Sasha, but I’m not sure and would have to double check
Predicted answer 3: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama and
       Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001.
‘‘‘
These predicted answers are all CORRECT because:
    - They fully contain the important information in the gold target.
    - They do not contain any information that contradicts the gold target.
    - Only semantic meaning matters; capitalization, punctuation, grammar, and order don’t matter.
    - Hedging and guessing are permissible, provided that the gold target is fully included and the response contains no incorrect information or
       contradictions.


The following are examples of INCORRECT predicted answers.
‘‘‘
Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: Malia.
Predicted answer 2: Malia, Sasha, and Susan.
Predicted answer 3: Barack Obama does not have any children.
Predicted answer 4: I think it’s either Malia and Sasha. Or it could be Malia and Jackie. Or it could be Joey and Malia.
Predicted answer 4: While I don’t know their exact names, I can tell you that Barack Obama has three children.
Predicted answer 5: It’s possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is
       that the correct answer?
Predicted answer 6: It may be the case that Obama’s child is named James. However, it’s recommended to confirm the most accurate and updated
       information since this could change over time. This model may not always reflect the most current information.
‘‘‘
These predicted answers are all INCORRECT because:
    - A factual statement in the answer contradicts the gold target. Incorrect statements that have some hedging (e.g., "it is possible that",
       "although i’m not sure, i think") are also considered incorrect.


The following are examples of NOT_ATTEMPTED predicted answers.
‘‘‘
Question: What are the names of Barack Obama’s children?
Gold target: Malia and Sasha
Predicted answer 1: I don’t know.
Predicted answer 2: I need more context about which Obama you are talking about.
Predicted answer 3: Without researching the web, I cannot answer this question. However, I can tell you that Barack Obama has two children.
Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I’m not sure about the other one.
‘‘‘
These predicted answers are all NOT_ATTEMPTED because:
    - The important information in the gold target is not included in the answer.
    - No statements in the answer contradict the gold target.


Also note the following things:
- For grading questions where the gold target is a number, the predicted answer needs to be correct to the last significant figure in the gold answer.
       For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k".
    - Predicted answers "120k", "124k", and 115k" are all CORRECT.
    - Predicted answers "100k" and "113k" are INCORRECT.
    - Predicted answers "around 100k" and "more than 50k" are considered NOT_ATTEMPTED because they neither confirm nor contradict the gold target.
- The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in
       the question.
    - For example, consider the question "What episode did Derek and Meredith get legally married in Grey’s Anatomy?" with gold target "Season 7,
       Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
    - For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer
       "San Francisco" would be considered CORRECT, even though it does not include "California".
    - Consider the question "What award did A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, &
       toxicity win at NAACL ’24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered CORRECT,
       because "award" is presumed in the question.
    - For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m". The predicted answer "1.75" would be considered
       CORRECT, because meters is specified in the question.
    - For the question "What is the name of Barack Obama’s wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be
       considered CORRECT, because the last name can be presumed.
- Do not punish for typos in people’s name if it’s clearly the same name.
    - For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong",
       "Hyungwon Chung", or "Hyun Won Chung".


Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT ATTEMPTED. Don’t apologize or correct yourself if there was a mistake; we are
       just trying to grade the answer.
‘‘‘
Question: {question}
Gold target: {target}
Predicted answer: {predicted_answer}
‘‘‘


Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: NOT_ATTEMPTED

Just return the letters "A", "B", or "C", with no text around it.
```

Figure 7: Prompt template used for the LLM-as-a-judge evaluator for ECLeKTic.

| Split | Exact-Match Accuracy (%) | LLM-as-a-Judge Accuracy (%) | Agreement |
|---|---|---|---|
| Source | 31.4 | 31.4 | 100.00% |
| Target | 23.1 | 23.2 | 99.94% |
| Overall | 23.8 | 23.9 | 99.95% |

Table 9: Comparison of Exact-Match and LLM-as-a-Judge accuracies. The near-perfect agreement (>99.9%) demonstrates the high reliability of our LLM-based evaluation protocol.

**Algorithm 1** Two-Stage Method for Model Clustering

**Require:** Set of models $M$, cutoff date $D_c$, scores $\mathcal{S} = \{(s_i, t_i)\}_{i=1}^{|M|}$
**Ensure:** Optimized clusters $C^*_{\text{old}}, C^*_{\text{new}}$

**Stage I: Temporal Initialization**
$C_{\text{old}} \leftarrow \{m \in M \mid \text{release}(m) < D_c\}$
$C_{\text{base_new}} \leftarrow \{m \in M \mid \text{release}(m) \geq D_c\}$
**Stage II: MSE Minimization**
min_MSE $\leftarrow \infty$
**for** each subset $S' \subseteq C_{\text{base_new}}$ where $|S'| \geq 3$ **do**
  $C_{\text{candidate_new}} \leftarrow S'$
  $C_{\text{candidate_old}} \leftarrow C_{\text{old}} \cup (C_{\text{base_new}} \setminus S')$
  $\text{WMSE}_{\text{total}} \leftarrow \sum_{C \in \{C_{\text{candidate_old}}, C_{\text{candidate_new}}\}} |C| \cdot \text{MSE}\left(\text{LinearFit}(C)\right)$
  **if** $\text{WMSE}_{\text{total}} <$ min_MSE **then**
    min_MSE $\leftarrow \text{WMSE}_{\text{total}}$
    $C^*_{\text{old}}, C^*_{\text{new}} \leftarrow C_{\text{candidate_old}}, C_{\text{candidate_new}}$
  **end if**
**end for**

**return** $C^*_{\text{old}}, C^*_{\text{new}}$

```
Judge whether the following ‘response‘ to ‘question‘ is correct or not based on the precise and unambiguous ‘correct_answer‘ below.

<question>
{question}
</question>

<response>
{predicted_answer}
</response>

<correct_answer>
{target}
</correct_answer>

The correct answer is always a number.  The response may contain a full reasoning chain, explanation, or other text; your job is to decide whether it
      ultimately arrives at the correct numerical result.

Your judgement must follow the format and criteria below:

<extracted_final_answer>
The final numeric answer extracted from the ‘response‘.  Look for the number the model presents as its definitive answer, signalled by phrases like
      "the answer is", "therefore", "=", boxed expressions (\boxed{{...}}), or a concluding statement.  Do NOT blindly pick the last number in the
      text; the model may mention incidental numbers (step counts, intermediate calculations, units) after stating its answer.  Put 'None' if there is
      no extractable numeric answer.
</extracted_final_answer>

<reasoning>
Explain why the extracted_final_answer is correct or incorrect based on ‘correct_answer‘.  Focus only on whether the numerical values match. Ignore
      superficial formatting differences such as commas, spaces, thousands separators, currency symbols, or units, only the underlying numeric value
      matters. Do not attempt to re-solve the problem.
</reasoning>

<correct>
Answer 'yes' if the extracted_final_answer matches the ‘correct_answer‘, or is within a small margin of error for numerical problems.  Answer 'no'
      otherwise, i.e. if the number differs, no number was found, or the model refused to answer.
</correct>
"""
```

Figure 8: Prompt template used for the LLM-as-a-judge evaluator for MGSMv2.

## A.3 Dual Fits

To investigate whether the scaling trends in Figure 1 undergo shifts across time, we employ a timeline-based partition on our models. As formalized in Algorithm 1, we utilize a two-stage clustering method to optimize these groupings.

In Stage I (Temporal Initialization), we establish base cohorts using a strict knowledge cutoff date of January 2025, grouping models into an older cohort and a modern base cohort (Table 10).

In Stage II (MSE Minimization), we refine these clusters to account for models in the modern cohort that have not exhibited characteristic progress breakthroughs (which are assymetric in temporal scale). We iteratively test subsets of the modern cohort, migrating candidate models to the older cluster. For each configuration, we fit independent linear regression lines and calculate the total weighted Mean Squared Error (WMSE). By selecting the partition that minimizes the combined MSE, we achieve a more robust dual-trajectory separation for our analysis. We plot our results in Figure 4 and observe that for MGSMv2, the dual linear fits successfully partition the models into two distinct temporal trajectories. Specifically, systems with a knowledge cutoff up to 2024 populate one baseline trajectory, while newer 2025 configurations shift to a separate, tighter linear regime. Furthermore, the median HAT score of these recent models significantly exceeds that of the older cohort.

## A.4 Accuracy on the Line for all target languages.

We separate our evaluation suite and analyze the joint distribution of source and target accuracy for every individual language across ECLeKTic, MMLU-ProX-Lite, and MGSMv2.

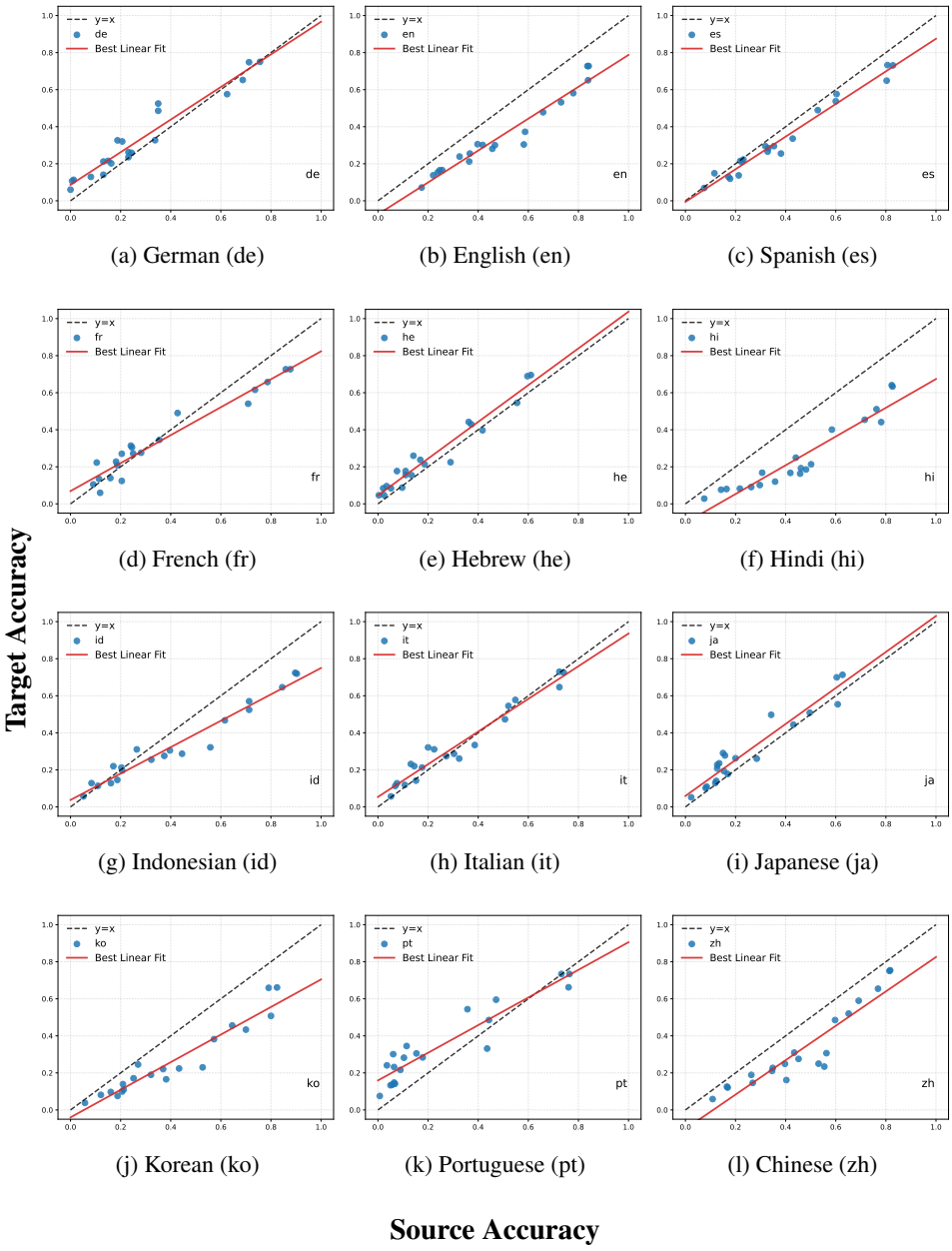
In this section, we recreate trends observed in Figure 1, separately for each target language. For ECLeKTic (Table 11) and MGSMv2 (Table 13), we report all respective target languages while for MMLU-ProX-Lite (Table 12), we only show 12 randomly chosen languages. The linear accuracy-on-the-line paradigm remains robust at the single-language level. This empirical behavior confirms that general performance improvements predictably propagate performance gains into individual target languages, providing strong empirical grounding for the instance-specific linear priors leveraged by our HAT score formulation in Section 4.1.

| Model Identifier | Month | Year |
|---|---|---|
| *Cohort 1: < 01-2025 Cutoff Date (△)* | | |
| GPT-OSS-20B | 06 | 2024 |
| GPT-OSS-120B | 06 | 2024 |
| GEMMA-3-1B | 08 | 2024 |
| GEMMA-3-4B | 08 | 2024 |
| GEMMA-3-12B | 08 | 2024 |
| GEMMA-3-27B | 08 | 2024 |
| QWEN-3-4B[4] | 11 | 2024 |
| QWEN-3-30B-A3B[4] | 11 | 2024 |
| *Cohort 2: ≥ 01-2025 Cutoff Date (□)* | | |
| GEMINI-2.5-FLASH | 01 | 2025 |
| GEMINI-2.5-FLASH-THINKING-OFF | 01 | 2025 |
| GEMINI-3-FLASH | 01 | 2025 |
| GEMINI-3-FLASH-LOW | 01 | 2025 |
| GEMINI-3-FLASH-MINIMAL | 01 | 2025 |
| GEMMA-4-E2B-IT | 01 | 2025 |
| GEMMA-4-E4B-IT | 01 | 2025 |
| GEMMA-4-26B-A4B-IT | 01 | 2025 |
| GEMMA-4-31B-IT | 01 | 2025 |
| CLAUDE-HAIKU-4.5 | 02 | 2025 |
| CLAUDE-SONNET-4.6 | 05 | 2025 |
| CLAUDE-OPUS-4.7 | 01 | 2026 |

Table 10: Partitioning of the models under study. Models are strictly categorized into pre- and post-January 2025 (**01-2025**) cohorts **based on their documented knowledge cutoff dates** (OpenAI et al., 2025; Google DeepMind, 2025b; Google, 2026a,b; Google DeepMind, 2026a; Anthropic, n.d.)]

[4]Since Alibaba does not appear to disclose an official Qwen3 knowledge-cutoff date, we treat November 2024 as a reasonable approximation: a secondary knowledge-cutoff tracker (Hillebrandt, 2026) lists both Qwen3-235B-A22B and Qwen3-32B (from the same family) with a November 2024 cutoff. However, a pre 01/2025 cutoff is highly likely given that the original Qwen3 series was publicly released on April 29, 2025, and that assertion is all we need for our analysis.

**ECLeKTic**




(a) German (de) (b) English (en) (c) Spanish (es)

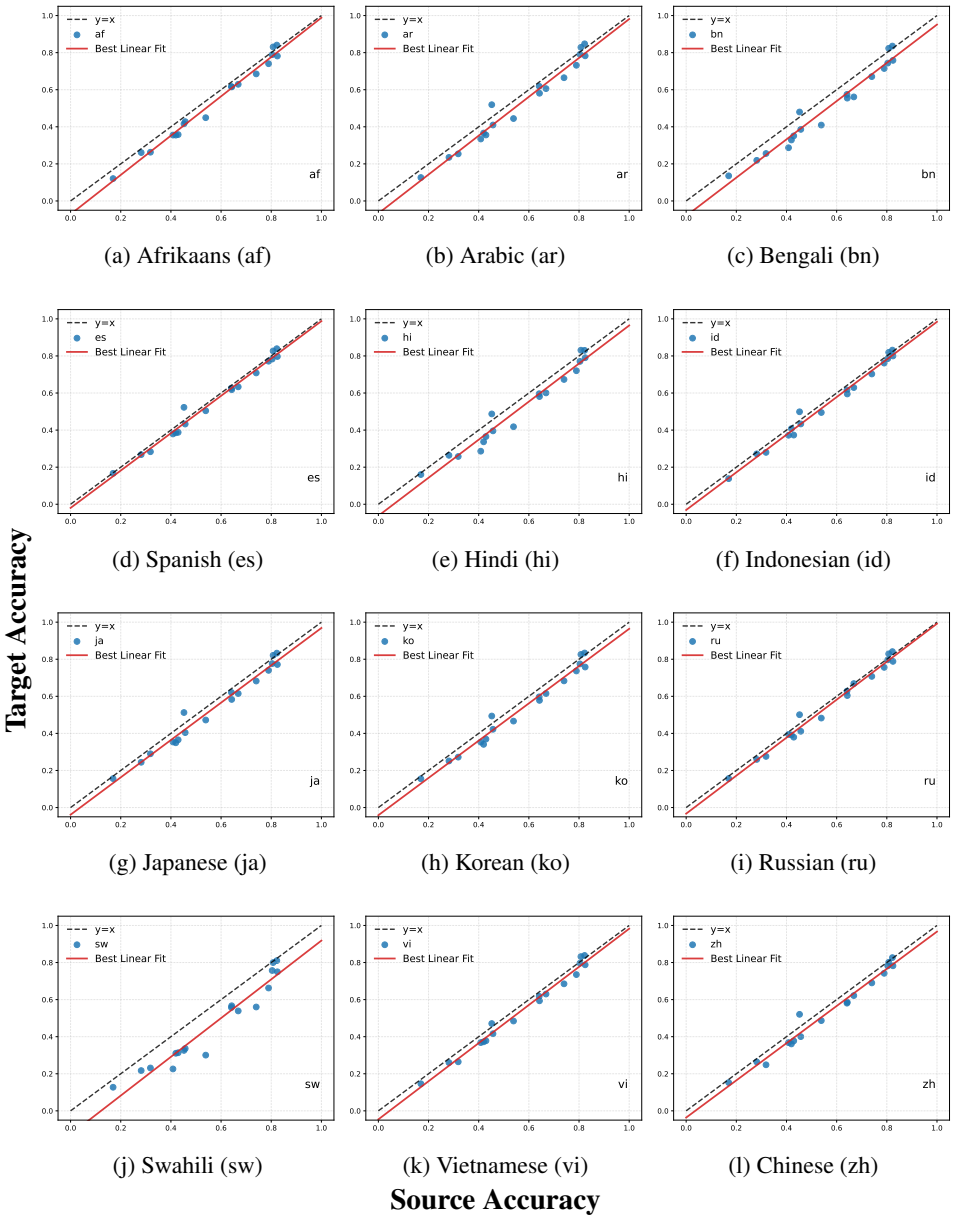
(d) French (fr) (e) Hebrew (he) (f) Hindi (hi)

(g) Indonesian (id) (h) Italian (it) (i) Japanese (ja)

(j) Korean (ko) (k) Portuguese (pt) (l) Chinese (zh)

Table 11: Detailed language-by-language breakdown of target accuracy versus source accuracy curves across the ECLeKTic dataset.

## MMLU-ProX-Lite




(a) Afrikaans (af) (b) Arabic (ar) (c) Bengali (bn)

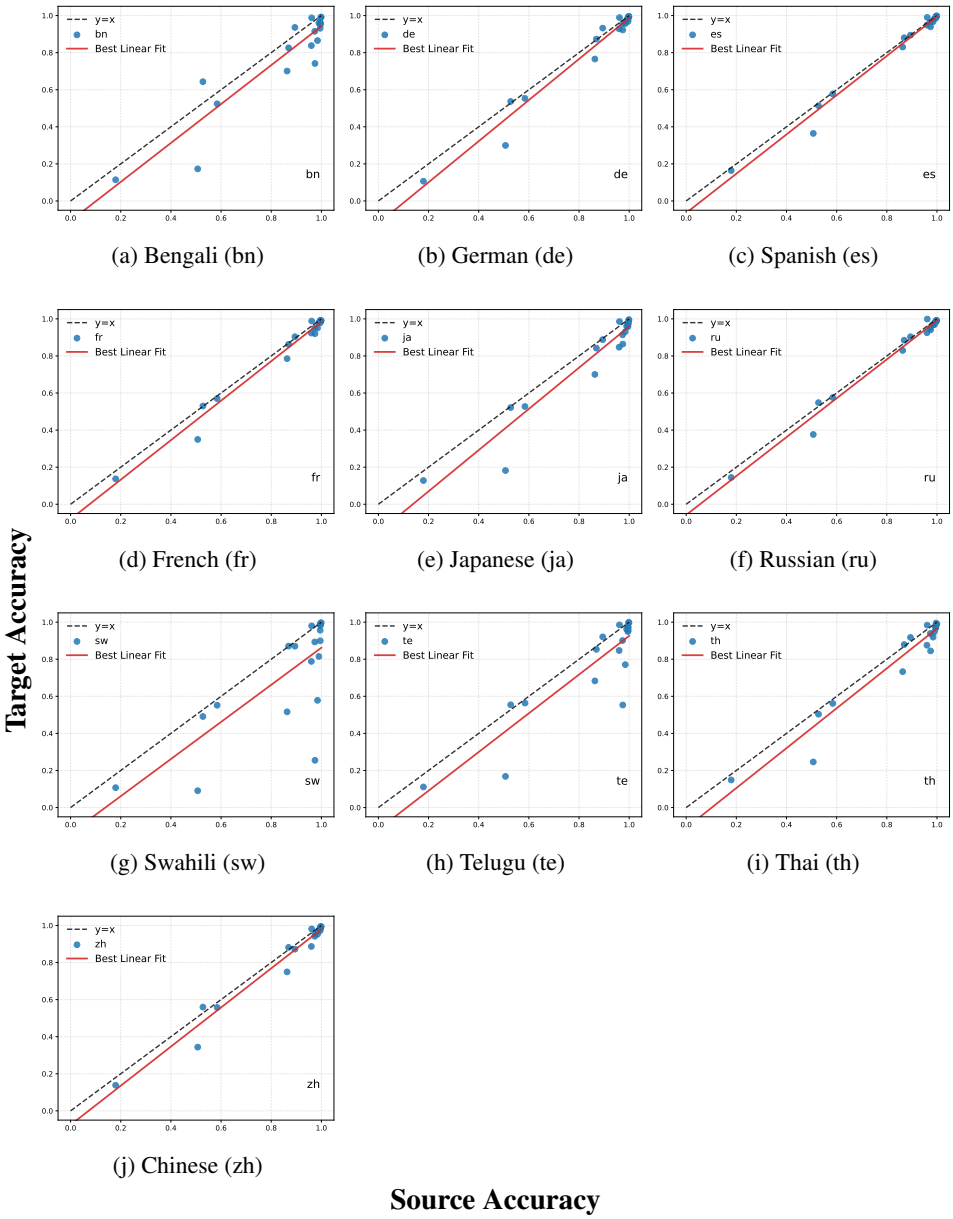
(d) Spanish (es) (e) Hindi (hi) (f) Indonesian (id)

(g) Japanese (ja) (h) Korean (ko) (i) Russian (ru)

(j) Swahili (sw) (k) Vietnamese (vi) (l) Chinese (zh)

Table 12: Detailed language-by-language breakdown of target accuracy versus source accuracy curves across the MMLU-ProX-Lite dataset.

## MGSMv2




(a) Bengali (bn) (b) German (de) (c) Spanish (es)

(d) French (fr) (e) Japanese (ja) (f) Russian (ru)

(g) Swahili (sw) (h) Telugu (te) (i) Thai (th)

(j) Chinese (zh)

Table 13: Detailed language-by-language breakdown of target accuracy versus source accuracy curves across the MGSMv2 dataset.

### A.5 Linear Argument

As introduced in Section 4.1 and illustrated in Figure 1 of the main text, our methodology relies on a linear prior for cross-lingual transfer. The purpose of this section is to empirically demonstrate that this linear assumption also translates robustly to the instance-specific level across different models.

Figure 9 visualizes this intra-model, instance-level transfer profile. The strong linear alignment observed here confirms that the linear relationship holds consistently at a granular, per-instance level, justifying its use as the foundational assumption for our HAT score formulation.

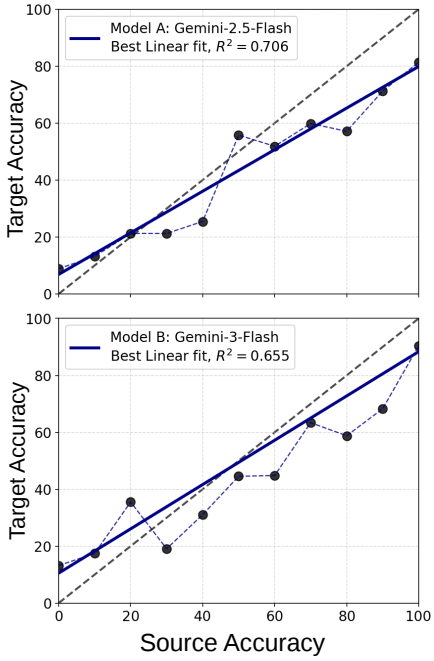



Figure 9: Intra-model, instance-level linear prior for the transfer profile.

### A.6 Comparison with the ECLeKTic Transfer Metric

We use the the definition of transfer score from the original study where it is defined for measuring XLT in ECLeKTic (Goldman et al., 2025). For a model $M$ evaluated on a source language $s$ and target language $t$, the transfer score is described as proportion of target questions answered correctly given they are answered correctly in source. Table 14 reports a comparison of Transfer scores and HAT scores.

For each question $q$, let $S \in [0, 1]$ be the source pass-rate and $T \in [0, 1]$ be the target pass-rate.

At a high level, the two metrics can be understood as solving structurally different optimization problems:

$$\text{HAT:} \quad \min_{m} \sum_{\text{all } q} \big(T(q) - m\, S(q)\big)^2 \tag{5}$$

$$\text{Transfer Score:} \quad \min_{c} \sum_{q:\, S(q)=1, T(q)\in\{0,1\}} \big(T(q) - c\big)^2 \tag{6}$$

Equation (5) uses every question; Equation (6) discards all questions with $S < 1$ and estimates a conditional mean on the remainder. This leads to three concrete advantages of HAT.

**1. All data contributes.** Questions answered with partial source pass-rates ($0 < S < 1$) carry genuine transfer signal: a model that transfers on questions it finds *difficult* in the source demonstrates stronger cross-lingual ability than one that transfers only on easy questions. The regression in (5) incorporates this signal; the Transfer Score ignores it.

**2. Robust for small models.** A weak model may answer very few questions correctly in the source ($|\{q : S(q) = 1\}|$ small), making the Transfer Score a high-variance estimate. HAT uses the full question set regardless of model strength, yielding stable estimates even for small models. Note that this weakness of the Transfer Score has already been laid out by authors proposed it (Goldman et al., 2025).

**3. Comparable across model scales.** The Transfer Score conditions on $\{q : S(q) = 1\}$, a set whose size and composition is *model-dependent*. A small model conditions on a narrow, easy subset; a large model conditions on a broader, harder-inclusive set. Their Transfer Scores are expectations over different populations, confounding transfer improvement with population shift. HAT avoids this: every model is scored on the same complete question set.

#### A.6.1 Empirical Divergence of HAT score from Transfer score

Table 14 reports both metrics side by side and reveals several cases where the Transfer Score paints a misleading picture.

| Model | ECLeKTic Transfer Score | | HAT Score | |
|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI |
| GEMINI-2.5-FLASH | 66.0 | [63.6, 68.4] | 81.7 | [78.6, 84.7] |
| GEMINI-2.5-FLASH THINKING-OFF | 44.9 | [42.5, 47.4] | 60.4 | [55.9, 64.8] |
| GEMINI-3-FLASH | 78.0 | [76.5, 79.6] | 84.6 | [82.1, 87.1] |
| GEMINI-3-FLASH LOW | 76.2 | [74.5, 77.9] | 84.0 | [81.6, 86.7] |
| GEMINI-3-FLASH MINIMAL | 68.0 | [66.1, 69.7] | 74.2 | [70.9, 77.5] |
| GEMMA-3-1B | 32.1 | [26.3, 38.0] | 48.2 | [34.9, 61.0] |
| GEMMA-3-4B | 45.8 | [41.4, 50.6] | 56.4 | [47.0, 65.9] |
| GEMMA-3-12B | 56.2 | [51.8, 60.5] | 64.1 | [56.8, 71.4] |
| GEMMA-3-27B | 58.8 | [55.1, 62.4] | 63.8 | [58.4, 69.6] |
| QWEN-3-4B | 38.8 | [34.3, 43.1] | 56.5 | [47.8, 64.9] |
| QWEN-3-30B-A3B | 48.9 | [45.3, 52.8] | 65.9 | [59.7, 72.6] |
| GPT-OSS-20B | 53.5 | [49.4, 57.5] | 80.1 | [74.7, 85.3] |
| GPT-OSS-120B | 61.0 | [58.0, 63.9] | 82.2 | [77.9, 86.1] |
| GEMMA-4-E2B-IT | 39.0 | [35.0, 43.2] | 51.8 | [42.4, 61.3] |
| GEMMA-4-E4B-IT | 43.9 | [40.0, 48.1] | 58.4 | [51.2, 65.7] |
| GEMMA-4-26B-A4B-IT | 55.2 | [51.8, 58.8] | 63.9 | [57.9, 69.6] |
| GEMMA-4-31B-IT | 55.3 | [51.9, 58.6] | 68.6 | [62.9, 74.1] |
| CLAUDE-HAIKU-4.5 | 51.1 | [48.2, 54.3] | 59.8 | [53.8, 65.7] |
| CLAUDE-SONNET-4.6 | 70.0 | [67.9, 72.2] | 74.7 | [70.9, 78.5] |
| CLAUDE-OPUS-4.7 | 67.0 | [64.9, 69.1] | 73.5 | [69.7, 77.2] |

Table 14: ECLeKTic Transfer and HAT Scores across models.

**Gemma scaling plateau** (illustrates advantage 1). The Transfer Score increases monotonically from `Gemma-3-1B` to `Gemma-3-27B` ($32.1 \rightarrow 58.8$), suggesting XLT improvement in transfer with scale. HAT, which also fits questions with partial pass-rates ($0<S<1$), tells a different story: scores rise from 1B to 12B ($48.2 \rightarrow 64.1$) but *plateau* at 27B ($63.8$). This could occur because of Transfer Score discarding the partial pass-rate signals, whereas HAT incorporates them to reveal the true underlying plateau.

**Gemini thinking ablation.** Disabling thinking in `Gemini-2.5-Flash` reduces the number of correctly answered source questions and degrades both metrics (Check Table 15 for performance in source). However, the Transfer Score drop (32% relative, $66.0 \rightarrow 44.9$) substantially exceeds the HAT drop (26% relative, $81.7 \rightarrow 60.4$). This exaggerated penalty demonstrates the Transfer Score's instability when it shrinks the pool of perfectly answered source questions (see Table 15)

**Claude Opus vs. Sonnet.** `Claude-Opus-4.7` is strictly more capable than `Claude-Sonnet-4.6` in the source language, yet receives a *lower* ECLeKTic score (67.0 vs. 70.0). HAT also ranks Sonnet slightly higher (74.7 vs. 73.5), but the gap narrows from 3.0 to 1.2 points. Because Opus is stronger in the source, it is evaluated on a broader, presumably harder to transfer question subset, than Sonnet. HAT is able to capture this.

### A.7 Model Performance in Source

We report the source performance of different models across the three datasets of interest, in the table 15.

### A.8 Average Thinking Token Counts

Table 16 presents the average thinking token counts for the source and target splits across the ECLeKTic, MGSM2, and MMLU-ProX-Lite datasets for the evaluated Gemini models.

| Model | ECLeKTic(%) | | MGSMv2 (%) | | MMLU-ProX-Lite (%) | |
|---|---|---|---|---|---|---|
| | Estimate | 95% CI | Estimate | 95% CI | Estimate | 95% CI |
| GEMINI-2.5-FLASH | 50.9 | [46.8, 55.0] | 99.5 | [99.0, 99.9] | 72.0 | [69.0, 75.2] |
| GEMINI-2.5-FLASH THINKING-OFF | 42.8 | [38.8, 47.1] | 58.4 | [52.6, 64.0] | 67.5 | [64.2, 70.8] |
| GEMINI-3-FLASH | 68.4 | [64.1, 72.5] | 99.8 | [99.4, 100.0] | 80.8 | [78.3, 83.3] |
| GEMINI-3-FLASH LOW | 69.1 | [65.0, 73.2] | 99.9 | [99.8, 100.0] | 82.2 | [79.6, 84.9] |
| GEMINI-3-FLASH MINIMAL | 65.7 | [61.4, 69.8] | 86.9 | [83.2, 90.6] | 80.4 | [77.4, 83.3] |
| GEMMA-3-1B | 5.4 | [3.7, 7.7] | 50.7 | [45.9, 55.6] | 17.0 | [14.1, 19.9] |
| GEMMA-3-4B | 11.5 | [8.9, 14.6] | 86.3 | [82.6, 89.6] | 28.2 | [24.7, 31.7] |
| GEMMA-3-12B | 17.3 | [14.0, 20.7] | 96.0 | [93.7, 97.8] | 42.0 | [38.1, 45.8] |
| GEMMA-3-27B | 26.4 | [22.2, 30.4] | 97.4 | [95.6, 98.9] | 45.7 | [41.8, 49.7] |
| QWEN-3-4B | 13.8 | [10.8, 16.9] | 97.4 | [95.9, 98.7] | 40.8 | [37.0, 44.5] |
| QWEN-3-30B-A3B | 19.4 | [16.2, 23.0] | 98.4 | [97.0, 99.5] | 53.9 | [50.2, 57.6] |
| GPT-OSS-20B | 19.7 | [16.5, 23.1] | 99.0 | [98.3, 99.6] | 74.0 | [71.0, 77.2] |
| GPT-OSS-120B | 27.9 | [24.3, 31.7] | 99.6 | [98.7, 100.0] | 79.0 | [76.0, 81.8] |
| GEMMA-4-E2B-IT | 11.9 | [9.1, 14.8] | 18.0 | [13.8, 22.3] | 31.8 | [28.5, 35.5] |
| GEMMA-4-E4B-IT | 15.7 | [12.7, 19.0] | 21.9 | [17.5, 26.6] | 42.9 | [39.0, 46.6] |
| GEMMA-4-26B-A4B-IT | 26.0 | [21.8, 29.8] | 99.5 | [98.6, 100.0] | 53.6 | [49.9, 57.5] |
| GEMMA-4-31B-IT | 28.2 | [23.8, 32.5] | 99.6 | [99.0, 100.0] | 64.1 | [60.4, 67.5] |
| CLAUDE-HAIKU-4.5 | 29.3 | [25.5, 33.7] | 52.8 | [46.4, 58.5] | 45.2 | [41.4, 49.1] |
| CLAUDE-SONNET-4.6 | 52.9 | [48.0, 57.8] | 96.1 | [93.7, 98.2] | 66.8 | [63.0, 70.7] |
| CLAUDE-OPUS-4.7 | 55.6 | [50.6, 60.0] | 89.4 | [85.6, 92.9] | 82.5 | [79.6, 85.2] |

Table 15: **Source Performance across ECLeKTic, MGSMv2 and MMLU-ProX-Lite.** Complete estimates and 95% confidence intervals for all tested model configurations.

| Dataset | Model | Split | Avg. Thinking Tokens | Proportion | English Fraction |
|---|---|---|---|---|---|
| ECLeKTic | GEMINI-2.5-FLASH | source | 276.1 | - | 0.952 |
| | GEMINI-2.5-FLASH | target | 314.9 | 1.15x | 0.965 |
| | GEMINI-3-FLASH | source | 264.5 | - | 0.951 |
| | GEMINI-3-FLASH | target | 315.9 | 1.19x | 0.959 |
| MGSM2 | GEMINI-2.5-FLASH | source | 269.0 | - | 0.999 |
| | GEMINI-2.5-FLASH | target | 273.3 | 1.01x | 0.992 |
| | GEMINI-3-FLASH | source | 283.2 | - | 0.999 |
| | GEMINI-3-FLASH | target | 298.0 | 1.05x | 0.990 |
| MMLU-ProX-Lite | GEMINI-2.5-FLASH | source | 1,118.4 | - | 0.991 |
| | GEMINI-2.5-FLASH | target | 1,258.2 | 1.13x | 0.970 |
| | GEMINI-3-FLASH | source | 689.3 | - | 0.989 |
| | GEMINI-3-FLASH | target | 675.6 | 0.98x | 0.980 |

Table 16: **Average token counts and English fraction.** Average tokens and the fraction of English tokens for source and target splits across ECLeKTic, MGSMv2, and MMLU-ProX-Lite datasets for Gemini 2.5 Flash and Gemini 3 Flash.