Lost in the Tower of Babel: The Adverse Effects of Incidental Multilingualism in LLMs

Anjishnu Mukherjee, Chutong Meng, and Antonios Anastasopoulos
Department of Computer Science
George Mason University
{amukher6,cmeng2,antonis}@gmu.edu

Abstract

This paper argues that contemporary multilingual NLP has converged on a fragile and misleading paradigm of *incidental multilingualism*. Today's LLMs appear multilingual largely because they are trained on massive, uneven web corpora, not because multilingual or multicultural competence has been treated as a core design objective. We contend that this paradigm systematically produces unequal, brittle, and opaque behavior across languages, with severe consequences in real-world and agentic deployments where models must reason, plan, and act across multiple linguistic contexts. We report a focused empirical study of two practical questions: which languages models self-report as supported and which languages they actually respond in across multilingual prompts. We additionally demonstrate how even a simple language-change attack can surface these failures and expose hidden assumptions about language in LLM-based systems. To address this, we call for a shift toward *multilingualism by design*: a research agenda that treats equitable multilingual performance, cultural grounding, and cross-lingual behavioral understanding as first-class goals in all aspects of the model pipeline.[1]

1 Introduction

Despite persistent and well-documented gaps, the success of large language models (LLMs) in multilingual settings is undeniable. Contemporary LLMs exhibit usable competence in over 100 languages and are increasingly deployed in real-world applications across the globe. This apparent breadth has led to a widespread perception that multilingual language modeling is a largely solved or at least steadily improving problem.

But the same amazing LLMs exhibit embarrassingly problematic limitations. One notable such limitation is their inability to consistently generate text in a user's desired language. Even if top-tier models such as GPT-4o score between 96 and 98% on the Language Confusion Benchmark (Marchisio et al., 2024), a system that receives more than 2 billion queries per day (like CHATGPT does), effectively still responds in the wrong language 40+ million times per day.

In this paper, we argue that this perception of success is somewhat misleading. Much of the recent progress in multilingual LLMs rests on shaky foundations that risk constraining future advances and obscuring systemic weaknesses. In particular, we contend that the dominant paradigm in multilingual LLM development relies on a set of implicit assumptions that deserve renewed scrutiny. Without revisiting these assumptions, further progress toward robust, equitable, and predictable multilingual systems may prove difficult.

We call multilingual capabilities of current LLMs that emerge largely as a side effect of other decisions, such as training on massive, heterogeneous web corpora, *incidental multilingualism*. We contrast it with *multilingualism by design*, where multilingual capability is treated as a primary design goal, and explicit objectives and success criteria across languages and

[1]Results, code and data: https://github.com/antonisa/llm-languages.

cultures shape data, modeling, evaluation, and deployment. Our experiments aim to diagnose behavioral symptoms and support gaps that might contradict even the training intent of LLM providers. To ground this argument in evidence relevant for deployment, we analyze a two-way support gap between self-declared support lists produced by models when asked directly, and observed non-refusal multilingual behavior under diverse task prompts.

This paper makes four contributions. First, we briefly situate our argument within the current status quo of multilingual language modeling (§2). Second, we present a focused investigation of language support claims versus observed behavior across current frontier models (§3). Third, we demonstrate how a simple language-change "attack" exposes failure modes of incidental multilingualism in practical LLM-powered agentic scenarios (§4). Finally, we outline a set of research directions and policy suggestions aimed at moving the field toward multilingualism by design (§5).

We aim for this paper to be neither an exhaustive survey nor an explicitly empirical one. Rather, it is a position piece grounded in concrete investigation and examples, intended to provoke discussion and re-examination of prevailing practices in multilingual LLM research.

2 The Status Quo in Multilingual NLP

Early massively multilingual modeling was driven by research in machine translation (e.g., Aharoni et al., 2019) and eventually hit the spotlight with encoder-only transformer language models such as multilingual BERT (mBERT) and XLM-R. mBERT (Devlin et al., 2019) was trained on Wikipedia in 100+ languages with a shared subword vocabulary, the first model that enabled zero-shot transfer across languages but also exhibiting strong performance disparities between high- and low-resource languages (Wu & Dredze, 2020). Generative LLMs such as GPT-*n* and various open-source families like LLaMA (Touvron et al., 2023) inherit this *incidental* multilingualism: they are trained on massive but heavily skewed web corpora, which gives them surprisingly broad language coverage but without explicit design for facilitating language transfer or equitable performance across languages (Doddapaneni et al., 2021; Qin et al., 2025). Below, we outline general observations that we believe hold for the vast majority of multilingual LLM research and development.

Uneven Performance and Evaluation Issues A consistent finding across models and tasks is that multilingual performance is highly imbalanced and, in many cases, "unfair". Accuracy can vary dramatically across languages even when tokenization and architecture are shared, with languages that are low-resource or typologically distant from English lagging behind. This was true for mBERT (Pires et al., 2019) and still is to this day for the most recent models (Song et al., 2023). Even worse, Ahuja et al. (2024) found robust indicators that several models are likely to be contaminated with multilingual evaluation benchmarks, making proper evaluation at scale almost impossible.

Much recent multilingual work focuses on evaluation and dataset construction to quantify these problems. Benchmarks such as XTREME (Ruder et al., 2021), XGLUE (Liang et al., 2020), FLORES (Goyal et al., 2022), MIRACL (Zhang et al., 2023), and multilingual safety benchmarks like XSafety (Wang et al., 2024) provide multi-task, many-language test suites that make cross-lingual gaps and safety issues more visible. Efforts like GlobalBench (Song et al., 2023), CultureBench (Chiu et al., 2025), or DialectBench (Faisal et al., 2024) aimed to organize benchmarking around specific goals (holistic LM evaluation, cultural appropriateness, and robustness to dialectal variation respectively).

The Aya initiative did push toward "multilingualism by design," at least on the data side: the Aya Dataset and Collection (Singh et al., 2024) constitute large-scale multilingual instruction corpora with deliberate coverage of 100+ languages, and Aya-23 (Aryabumi et al., 2024) traded breadth for depth to improve quality in 23 languages, releasing open-weight models and detailed multilingual evaluation to track progress (Cohere For AI, 2024).

In parallel, data- and tokenizer-centric efforts aim to mitigate structural disadvantages. The argument can be made that byte- or character-level models such as ByT5 (Xue et al., 2022) remove language-specific subword vocabularies and show competitive performance across

scripts, with follow-up work demonstrating that tokenization decisions materially affect low-resource and morphologically rich languages (Edman et al., 2024; Dang et al., 2025). More recent efforts even attempt to go tokenization-free by converting text inputs to images, in the PIXEL model series (Kesen et al., 2025; Tatariya et al., 2024).

Despite this intense activity around benchmarks, data curation, and tokenization, there is comparatively less work on *understanding* what massively multilingual (200+ language) models actually know and how languages interact internally. Unfortunately, properly understanding such complicated systems is highly computationally expensive: Malkin et al. (2022) rely on approximations from bilingual modeling, while Faisal & Anastasopoulos (2024) similarly relied on bi- and trilingual adapters on top of pre-trained multilingual models. Surveys emphasize that most analyses still focus on surface-level performance or a small subset of languages, leaving open questions about latent language clustering, cross-lingual interference, knowledge sharing, and how cultural and moral biases propagate across languages in very large models (Doddapaneni et al., 2021; Qin et al., 2025; Hämmerl et al., 2022; Jain & Chopra, 2026).

A telling sign of the field's current state, and of why our paper is still very much needed, is that we are still restating critiques made more than a decade ago. Bender (2009) warned against treating systems as language-independent simply because they are linguistically naïve. Bender (2011) further argued that the field was neither systematically evaluating language-independence nor building it with sufficient linguistic sophistication. Later work revisited the same structural problem from different angles: popular cross-lingual proxy evaluations can misrepresent downstream generalization (Glavaš et al., 2019), the English-centric allocation of research effort reproduces inequality (Søgaard, 2022), and much of NLP still defaults to a narrow "square one" paradigm centered on English and accuracy alone (Ruder et al., 2022). That these critiques remain salient in the LLM era suggests that multilinguality has too often been treated as an emergent byproduct of scale rather than as an explicit design objective.

In this sense, the field is in a transitional phase: we can increasingly measure and partially patch inequities in multilingual LLMs. And this is uncontroversially a crucial, necessary first step. But fully explaining how 100-200+ languages co-exist and shape one another inside these models in terms of both representations and socio-cultural behavior remains an important open research frontier.

Consequences in Real-World Settings The reliance on *incidental multilingualism* has direct and often problematic consequences for real-world deployment of LLMs. In principle, multilingual systems deployed in domains such as education, healthcare, public services, or customer support should offer comparable reliability, safety, and calibration across languages. In practice, however, performance and safety lack data availability and research attention. Consequently, what is often presented as a single global model behaves like a collection of language-conditioned systems with divergent capabilities and failure modes.

These disparities are particularly pronounced in safety and cultural adequacy. LLMs generate significantly more unsafe responses for non-English prompts (Wang et al., 2024), and multilingual jailbreaks exploiting low-resource languages are easier to construct and harder to defend against (Shen et al., 2024). At the same time, multilingual models are not inherently multicultural (Adilazuarda et al., 2024). They frequently encode Anglocentric norms even when operating in other languages, leading to culturally misaligned or inappropriate behavior. LLM safety evaluation and mitigation efforts overwhelmingly focus on English, leaving safety guarantees in other languages underdeveloped (Yong et al., 2025). In deployment, this translates into higher risks precisely for the most vulnerable users with the least institutional oversight or recourse.

The growing use of LLM-powered *agentic systems* amplifies these concerns. Such systems must interpret instructions, retrieve information, and act across multiple linguistic contexts, yet even monolingual agent reliability remains an open research problem (Cheng et al., 2024). Introducing multilinguality creates additional failure modes (Chen et al., 2024). Each language boundary potentially enables semantic drift, misalignment, or unsafe actions, underscoring the risks of deploying agentic systems built on incidental multilinguality.

From Ideal Design to Current Practice Taken together, the previous notes highlight a growing gap between an "ideal" multilingual-by-design ecosystem and current practice. In the ideal picture, models and agents would be trained and evaluated with explicit language- and culture-aware objectives; safety mitigations, datasets, and benchmarks would be co-developed across languages; and real-world systems would guarantee comparable reliability and protections regardless of users' background. By contrast, most deployed LLMs today still rely on English-centered safety tooling. This tension is particularly acute for agentic systems, where multilingual failures can propagate into actions rather than remaining as text. Bridging this gap will require a shift toward treating multilingualism and multicultural competence as central design goals.

3 What Languages Do LLMs Speak?

In practice, "language support" is not a single property. We offer here at least three notions, which can diverge sharply: (1) the languages a model claims to support when asked directly, (2) the languages in which it can produce text that is actually identified as the target language, and (3) the languages in which it can perform nontrivial downstream tasks reliably.[2] When these notions disagree, both users and deployers lose a stable operational definition of support. This distinction matters in high-stakes settings such as education, public services, healthcare, and legal assistance, where deployment decisions are often justified by broad multilingual claims rather than by task-specific evidence. Our question in this section is therefore deliberately operational: when we move from self-description to behavior, what languages do current frontier LLMs actually "speak"?

3.1 Experimental Design

We evaluate five frontier LLMs: GPT-5.4, CLAUDE OPUS 4.6, GEMINI 3.1 FLASH-LITE, LLAMA 4 MAVERICK, and QWEN 3.5 397B. Our goal is to measure both what languages a model claims to support (*declared support*) and how it behaves on downstream tasks in those languages (*behavior*).

Declared support We first study how models describe their own multilingual coverage.

(1) *Elicitation.* Each model receives 20 English prompt variants asking for a JSON list of supported languages and ISO codes (Figure 1).

(2) *Verification.* Since provider documentation for these models does not give a complete language-by-language ground-truth list, we cannot score these support lists directly. Instead, for each of the 20 elicitation runs, we ask the model to generate a sentence in every language it claimed in that specific run. Thus, verification is repeated once per claimed language per prompt variant. We then pass each generated sentence through GLOTLID (Kargaran et al., 2023) to confirm whether the detected language matches the claimed one. Figure 1 reports intersection (I) and union (U) bounds over both the raw language sets and the confirmed language sets across the 20 prompts.

(3) *Language support awareness.* We next evaluate two additional tasks on each model's LangID-confirmed supported-language union bound U. First, we provide the model with a sentence in one of these confirmed languages and ask it to identify the language code (*lang_id*). Second, we provide the name of that language and ask the model to answer *yes* or *no* when asked whether it can write or speak it (*language_self_report*). If models were well calibrated on their own confirmed support frontier, they should perform near-perfectly on both tasks. We score *lang_id* by exact correctness of the predicted language code, and *language_self_report* by whether the model answers *yes* for languages in its own LangID-confirmed support frontier U.

[2] Of course, there's also the notion of "official" support by way of listing languages in model cards or release notes. But no major provider provides such a list.

Behavior We next evaluate multilingual behavior on downstream tasks (Figure 2). These task families use a shared 203-language collection: English along with the 202 non-English languages in a FLORES/NLLB-backed translatable set. Non-English prompt translations are generated with NLLB-200 3.3B (NLLB Team et al., 2022). Appendix B audits NLLB prompt quality with COMETKiwi (Rei et al., 2022) and GlotLID on known-language reference samples.

(1) *Short story generation.* We ask the model to generate a culturally relevant short story of roughly 4-6 sentences in the target language.

(2) *Teacher conversation.* We assign the user the persona of a language learner and ask the model to hold a short pedagogical conversation in the target language.

For both of these tasks, we evaluate whether the output is in the requested language using GLOTLID.

(3) *Translation to/from English.* We ask models either to translate text into English or from English into the target language while preserving meaning and tone. We evaluate these tasks in two ways: target-language retention is measured with GLOTLID, and translation quality is measured with CHRF++ (Popović, 2017).

(4) *Code generation.* We describe a programming task in natural language and ask the model to generate Python code. We evaluate this task against MHUMANEVAL (Raihan et al., 2025). We use *parseability*, indicating whether the generated code can be parsed into an abstract syntax tree, and *functionality*, indicating whether the generated code passes the benchmark unit tests, as our metrics.

(5) *Long-form generation.* Last, we evaluate open-ended generation using POLYWRITE (Ji et al., 2025). Although this benchmark is available for 240 languages, only 141 overlap with the FLORES collection used in our other tasks (and we only report results on this subset).

3.2 Results and Discussion

For our results below, we report performance grouped by the language resource classes of Joshi et al. (2020), which classified languages according to their resource availability. The downstream task families are aggregated over repeated runs of the current evaluation pipeline, and we summarize model performance with the task-specific metrics above. We also explicitly permit models to refuse when they cannot complete a task reliably. In practice, however, the dominant failure mode is usually incorrect generation rather than refusal.

Declared support Figure 1 and Table 3 show that the multilingual support frontier exposed by these models is neither stable nor well calibrated.

(1) *Elicitation.* Prompting the same support-list question in slightly different paraphrases, yields a wide spread of claimed support frontiers. Raw support unions range from just 18 languages for GEMINI 3.1 FLASH-LITE to 187 for both GPT-5.4 and CLAUDE OPUS 4.6, with LLAMA 4 MAVERICK reaching 183. Even within a single model, counts vary sharply across paraphrases, for example LLAMA 4 MAVERICK ranges from 1 to 183.

(2) *Verification.* After verifying the sentences generated for each language the models claim to support, the union U decreases for every model, more for some and less for others. Raw self-declared support is thus consistently too optimistic, and verification makes the behavioral frontier substantially more conservative. Low intersection I is an even bigger problem. Even after verification, confirmed intersections remain small relative to the union upper bound. This means many languages are only intermittently acknowledged as supported across paraphrases. LLAMA 4 MAVERICK suffers the most from this phenomenon, as its confirmed union is broad ($U = 134$), yet only one language (English) survives across all 20 prompt variants and reruns.

> **Takeaway 1:** Declared multilingual support is unstable under simple prompt variation, and LangID verification reduces the set of languages models can plausibly claim to support.

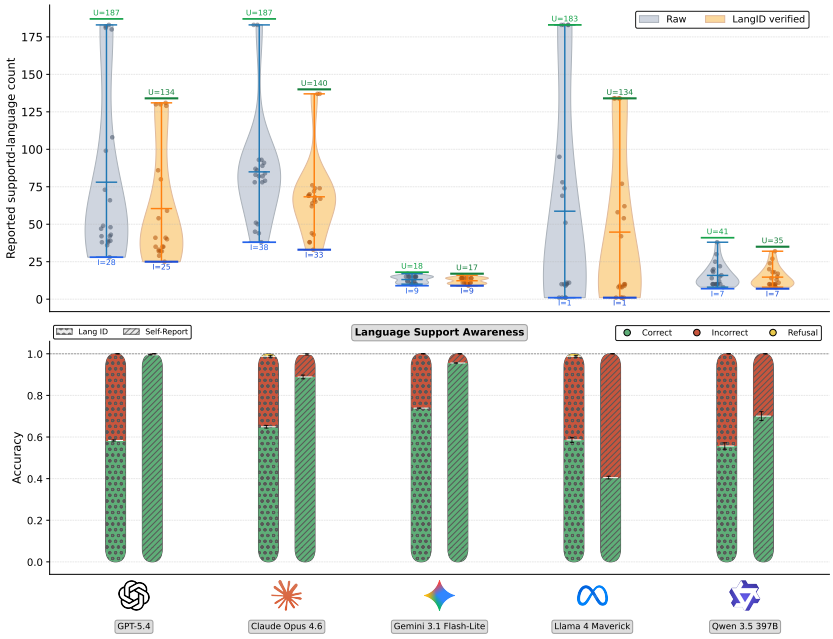

Figure 1: Language support claims are highly prompt sensitive (top, *declared support*) and are not well calibrated on each model's confirmed support frontier (bottom, *behavior*). The upper panel shows how much the set of supposedly supported languages changes across 20 English paraphrases of the same question. The lower panel evaluates language identification and direct yes/no self-report only on languages that are LangID verified for that model.

(3) *Language support awareness.* Table 3 shows that the *language_self_report* task is generally easier for models than target language identification (*lang_id*), implying that these models claim support in languages where they cannot reliably identify text. These results support our call for much stronger provider transparency: models should expose versioned language-support frontiers and explicitly decline languages that fall outside them.

Behavior Figures 2 and 3, along with Table 4, show that multilingual behavior is strongly task-dependent, and that incorrect generation is far more common than explicit refusal.

(1) *Short story generation* and (2) *teacher conversation.* Short-form target-language generation is comparatively close across models. Most are generated in the requested language, but the open-source models are the only ones that regularly refuse requests.

(3) *Translation to/from English.* GPT-5.4 leads both translation CHRF++ directions (35.44 EN→X and 45.46 X→EN). Translation quality and generating in the requested language measure different abilities, and the results show that generating in the correct language does not always imply a high-quality translation.

(4) *Code generation.* CLAUDE OPUS 4.6 leads code functional correctness (0.855), while parseable code is generally easier to generate for all models than functionally correct code. This gap shows that models can often exploit strong Python priors to produce syntactically valid code even without fully capturing the task requirements.

For these four downstream tasks, non-refusal rates typically remain above 0.94 and often above 0.98, while correctness is much lower. So where models should arguably abstain, they instead produce incorrect language output, poor translations, or non-functional code. This matters operationally because the potential user sees an apparently cooperative model rather than a transparent admission of weak support.

> **Takeaway 2:** Across downstream multilingual tasks, the dominant failure mode is usually incorrect generation rather than explicit refusal.

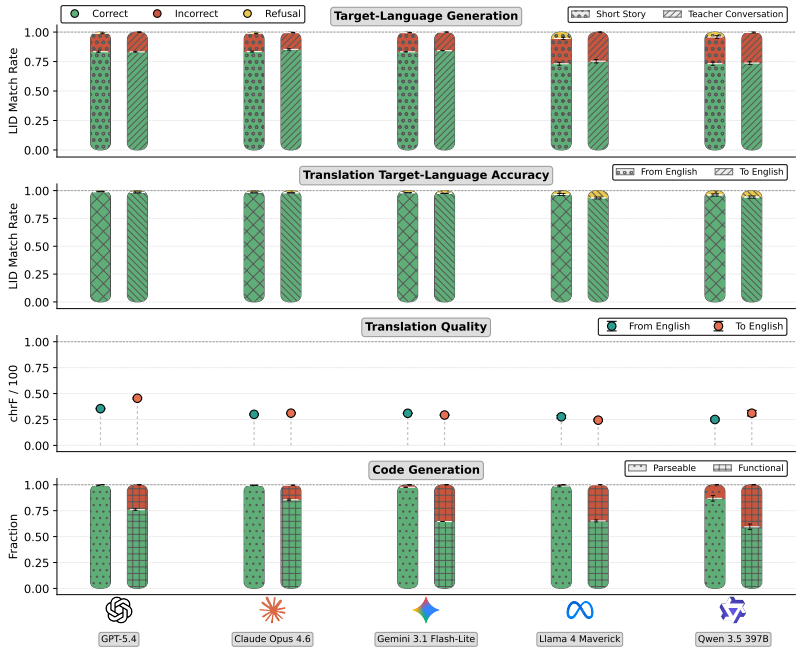

Figure 2: Multilingual task performance on 203-languages across 4 tasks - Short stories and pedagogical dialogues test whether models can generate in the requested language, the translation panels separate ability to generate in target-language from translation quality measured with CHRF++, and the code generation panel illustrates the difference between parseable Python and functionally correct code on multilingual MHUMANEVAL prompts.

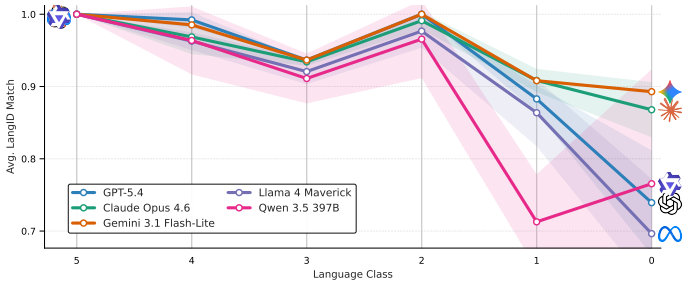

Figure 3: Target-language retention in long-form writing across 141 languages, grouped by six resource classes (Joshi et al., 2020). Class 0 contains languages with the least language resources and class 5 the most. The hardest classes widen the gaps between models.

(5) *Long-form generation.* Figure 3 shows clear differences across models as we progress from high-resource languages to lower resourced buckets. All models saturate class 5 and remain close in classes 2-4, but have distinctly separate performance for classes 0-1. So the most meaningful differences between models appear not in high-resource languages, where all do well, but in low-resource settings where replying in the requested language is much less reliable and broad claims about coverage fail to predict actual usability.

> **Takeaway 3:** The hardest low-resource languages remain the clearest separator between models, and broad support claims are least predictive exactly where multilingual robustness matters most.

Discussion These empirical results argue against treating multilingual support as a binary model attribute. At minimum, reporting should separate *declared support*, *stable support*, and *task-conditional behavioral support*. Otherwise a claim such as "supports 100+ languages" collapses over phenomena that are empirically distinct and operationally consequential.

Our comparison of frontier models also suggests that there is no single multilingual ordering. GPT-5.4 leads in translation, CLAUDE OPUS 4.6 in multilingual code correctness, GEMINI 3.1 FLASH-LITE in target-language generation for the hardest POLYWRITE buckets, while LLAMA 4 MAVERICK and QWEN 3.5 397B are competitive on some mid-resource slices despite weaker support-list stability. Multilingualism is therefore not a monolithic capability that can be summarized by one headline number. It is a structured capability profile, with different models tracing out different frontiers across support claims, translation, language retention, and long-tail generation.

4 The Tower of Babel Problem

Recent work increasingly explores *collaborative* and *agentic* LLM systems, in which multiple language-model-based agents interact, and jointly produce outputs. A representative example is DYLAN (Liu et al., 2024), where several agents (instantiated from the same underlying LLM) independently propose solutions, evaluate one another's responses, and are dynamically selected to collaborate through iterative discussion rounds. Such systems implicitly assume that communication between agents is semantically stable and that linguistic variation does not fundamentally alter reasoning, coordination, or task performance.

This assumption though is fragile in the presence of incidental multilingualism. Modern LLMs exhibit language-dependent behavior, performance, and alignment properties. In some cases, such variation may be appropriate, for example, when agents are expected to adapt to linguistic or cultural context. However, agentic systems that operate across languages are vulnerable to failures that arise purely from linguistic heterogeneity rather than task complexity. We refer to such failures as the **Tower of Babel (ToB) problem**: a scenario in which otherwise identical agents are induced intentionally or accidentally to communicate in different languages, thereby degrading coordination and task success.

It would be straightforward for an adversary to devise a ToB attack, exploiting the fact that LLM behavior is not invariant across languages. The intuition is that in practical deployments, agents may interact with heterogeneous external data sources, user inputs, or retrieved documents in multiple languages. This creates a natural surface for accidental or adversarial language shifts through multilingual web content, translation artifacts, or data poisoning or prompt injections that nudges agents to respond in specific languages. Under incidental multilingualism, there is no guarantee that agents operating in different languages will reason, evaluate, or communicate in mutually compatible ways, even when they share the same underlying model.

	all English	ToB attack
MMLU accuracy	75.0	67.5

Figure 4: Task performance degrades substantially when agents operate multilingually.

Illustrative proof of concept As a demonstration, we simulate the ToB issue in a simplified collaborative-agent setting. Similar to Liu et al. (2024), we instantiate seven role-based agents (`Economist`, `Doctor`, `Lawyer`, `Mathematician`, `Psychologist`, `Programmer`, `Historian`) and evaluate them on MMLU-REDUX (Gema et al., 2025). We compare two conditions: (a) a monolingual baseline where all agents communicate in English, and (b) a multilingual condition where each agent is tricked into operating in a different language (English, Mandarin Chinese, isiZulu, Greek, Hungarian, Bengali, Quechua, Turkish, or Romanian).[3] Aside from language, the setup is identical across conditions.

Second-order effects The DYLAN framework relies on *inter-agent evaluation*: each agent scores the initial responses of the others, and these scores determine which agents are selected to participate in the collaborative refinement stage. In principle, these evaluations should reflect the substance, correctness, and usefulness of each response, independent of the language in which it is expressed.

[3]Languages are chosen mostly for geographic and typological diversity.

But, we observe language-dependent differences in these scores. For example, Greek-speaking agents consistently assign Turkish-speaking agents scores approximately one standard deviation lower than those assigned to agents operating in Western European languages (and vice versa). Similar patterns emerge (to a smaller extent) between the Romanian and Hungarian agents.[4] Perhaps unsurprisingly, agents operating in Quechua and isiZulu receive substantially lower scores from nearly all other agents, regardless of task content. While these descriptive patterns are anecdotal and some of this effect may be attributable to limited or uneven model fluency in these languages, the key observation is that language alone exerts a strong influence on agent selection dynamics.

Implications The performance drop in Figure 4 occurs without model or task changes and is consistent with language-induced variation in agent behavior, although our design does not explicitly separate individual comprehension from coordination effects. While simple, the result illustrates how incidental multilingualism may undermine *performance* and *coordination* in agentic LLM systems. Language-dependent biases in inter-agent evaluation suggest a possible second-order failure mode: agents operating in certain languages are systematically devalued regardless of the quality of their contributions. In real-world deployments, this risks reproducing linguistic and cultural hierarchies inside ostensibly neutral AI systems, where language choice implicitly shapes authority, trust, and influence.

5 Toward Multilingualism by Design: Some Concrete Directions

The reader is hopefully convinced by now that research must treat multilingual capability as an explicit design objective across the full LLM pipeline. Below we offer a set of concrete directions that go beyond the usual calls for curated data selection, balanced data mixtures, or calls for extensive multilingual evaluation and culturally grounded benchmarks.[5]

At the core of multilingual language modeling lies a fundamental tension between *language-agnostic representation learning* and *language- and culture-specific behavior*. Language agnosticism enables cross-lingual transfer and parameter sharing. But language- and culture-specificity is essential for appropriate reasoning, generation, and interaction in real-world contexts. Strongly shared representations such as those induced by joint subword vocabularies and fully shared transformer parameters are the current paradigm, but they also risk collapsing meaningful linguistic, cultural, and pragmatic distinctions. In many settings, it is absolutely fine if "LLMs reason in English" (Etxaniz et al., 2024), but in many others it would be catastrophic. In our view, the translate-test and in general translation-based approaches (Liu et al., 2025; Ebrahimi & von der Wense, 2024) are problematic fixes to a persistent problem that can never become an appropriate general solution regardless of how good translation capabilities become.

Conversely, introducing language-specific components or objectives can preserve local variation, but may reduce transfer efficiency and complicate scaling, a trade-off that remains poorly understood. Somewhat-forgotten work from the encoder LLMs era on explicit alignment objectives (Hu et al., 2021) suggests that this trade-off is not binary, especially if put into an envisioned context of more carefully curated multilingual data at scale such as the Aya efforts (Singh et al., 2024). Careful design choices should improve cross-lingual alignment without erasing language-specific signals, and even drive overall performance up. Or at a minimum, the system design should be based on a pre-specified target along this tension continuum.

However, practitioners still lack principled guidance on how and where this balance should be managed. Different interventions at the level of tokenization, intermediate representations, training objectives, decoding, or post-tuning have been explored. These contributions are certainly steps in the correct direction, but substantial progress requires that they are perhaps employed in combination instead of as possible one-off, often post-hoc fixes.

[4]To be fair, our choice of several Balkan languages might lead to exaggerated effects due to the region's history. But it certainly supports our point about unintended effects.

[5]Not because these are not needed -*au contraire!*- but because they do not challenge the fundamental issues this paper aims to highlight.

Last, we argue that more work is needed in properly understanding multilinguality as reflected in LLMs. Our models appear to have both language-invariant components (Foroutan et al., 2022) and language-specific subnetworks (Tang et al., 2024), and both of these are crucial for different aspects of performance. This is perhaps unsurprising, but the mechanisms through which such components arise internally, how they might be shaped by cross-language interactions (be that positive transfer or negative interference), or how to identify them so as to study them, are all entirely open questions.

In the more specific context of collaborative LLM-powered *agents*, we note that current setups are fundamentally constrained by the requirement to communicate and coordinate exclusively through human language. Human languages are a medium shaped by human communicative needs and embodied assumptions rather than by the requirements of machine reasoning or coordination. While human languages provide an accessible interface for interaction, they may also introduce ambiguity or cultural biases when used as the sole channel for inter-agent communication. This raises the question of whether agentic systems should rely entirely on natural language internally, or whether alternative, non-linguistic or hybrid communication representations designed for stability, precision, or cross-lingual invariance could better support robust coordination in multilingual settings. Early work explored machine-to-machine communication (Tebbifakhr et al., 2019), as has the field of emergent communication (Cao et al., 2018), but such approaches have not found fertile ground in the current practical LLM-dominated environment. Exploring such representations might be a promising direction for decoupling agent reasoning and coordination from the incidental, often unwanted baggage that human languages carry.

Provider transparency and accountability We also advocate for explicit multilingual accountability standards at release time. Providers should publish versioned supported-language lists, together with language-level evaluation cards that report both quality and safety metrics rather than broad headline claims of multilingual support. They should further document language-conditioned refusal policies, making clear when support is considered inadequate and when the model is expected to abstain instead of improvising.

When a model cannot reliably support a language, it should transparently refuse rather than produce overconfident low-quality outputs. Such refusals should be paired with safer fallback guidance: for example, suggesting a better-supported language, requesting clarification, or directing the user toward a human escalation path where appropriate. This is not a complete substitute for multilingual capability, but it is a concrete harm-reduction mechanism. In the current paradigm, users are often exposed to hidden failure behind fluent-looking text; refusal-by-design makes those capability boundaries explicit and aligns model behavior more closely with what the system can actually do well.

It is worth pointing out that this is a question requiring hard language-specific decisions. The distinctions between languages are never clear cut, and this translates to different results for different tasks. For example, a multilingual LLM that is extremely competent in Spanish and Portuguese might be adequately capable of *understanding* and executing an instruction given in Mirandese or Asturian, but woefully inaccurate in actually *generating* fluent Mirandese text output. Thus, careful consideration should be given into whether a language like Mirandese should be included in the language support or whether any refusals should be task-specific. We do have a concrete answer to how one should balance the potential harms from the potential utility a language speaker might receive, but LLM providers should at least be transparent about the decisions that guide their products.

6 Conclusion

Modern LLMs exhibit broad multilingual behavior, but our findings show that official support claims, self-declared support lists, and observed multilingual behavior can diverge sharply. This mismatch is a direct symptom of incidental multilingualism: language capability emerges, but is not specified as a robust contract. Moving toward multilingualism by design requires explicit support definitions, multilingual evaluation transparency, and safe refusal behavior when support is inadequate.

Acknowledgments

We are grateful to the anonymous reviewers and area chairs from both COLM and ARR whose feedback improved the paper. Part of this work (the question of which languages frontier LLMs actually support) was the result of thoughtful discussions with Yulia Tsvetkov, Sunayana Sitaram, and Monojit Choudhury. This work is financially supported by the National Science Foundation under CAREER award 2439202.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in LLMs: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15763–15784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.882. URL https://aclanthology.org/2024.emnlp-main.882/.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2598–2637, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.143. URL https://aclanthology.org/2024.naacl-long.143/.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. Aya 23: Open weight releases to further multilingual progress. *arXiv:2405.15032*, 2024. URL https://arxiv.org/abs/2405.15032.

Emily M. Bender. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pp. 26–32, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-0106.

Emily M. Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011. doi: 10.33011/lilt.v6i.1239. URL https://journals.colorado.edu/index.php/lilt/article/view/1239.

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. Emergent communication through negotiation, 2018. URL https://arxiv.org/abs/1804.03980.

Chung-Chi Chen, Hiroya Takamura, Ichiro Kobayashi, and Yusuke Miyao. The impact of language on arithmetic proficiency: A multilingual investigation with cross-agent checking computation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 631–637, Mexico

City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.53. URL https://aclanthology.org/2024.naacl-short.53/.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. Exploring large language model based intelligent agents: Definitions, methods, and prospects, 2024. URL https://arxiv.org/abs/2401.03428.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalBench: A robust, diverse and challenging benchmark for measuring LMs' cultural knowledge through human-AI red-teaming. 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. 25663–25701, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1247. URL https://aclanthology.org/2025.acl-long.1247/.

Cohere For AI. The ai language gap: Considerations on the multilingual capabilities of ai language models. Policy primer, Cohere For AI, 2024. URL https://cohere.com/research/papers/the-ai-language-gap.pdf.

Thao Anh Dang, Limor Raviv, and Lukas Galke. Tokenization and morphology in multilingual language models: A comparative analysis of mT5 and ByT5. In Mourad Abbas, Tariq Yousef, and Lukas Galke (eds.), *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pp. 242–257, Southern Denmark University, Odense, Denmark, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.icnlsp-1.24/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. A primer on pretrained multilingual language models, 2021. URL https://arxiv.org/abs/2107.00676.

Abteen Ebrahimi and Katharina von der Wense. Zero-shot vs. translation-based cross-lingual transfer: The case of lexical gaps. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 443–458, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.37. URL https://aclanthology.org/2024.naacl-short.37/.

Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, and Arianna Bisazza. Are character-level translations worth the wait? comparing ByT5 and mT5 for machine translation. *Transactions of the Association for Computational Linguistics*, 12:392–410, 2024. doi: 10.1162/tacl_a_00651. URL https://aclanthology.org/2024.tacl-1.22/.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in English? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–564, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.46. URL https://aclanthology.org/2024.naacl-short.46/.

Fahim Faisal and Antonios Anastasopoulos. An efficient approach for studying cross-lingual transfer in multilingual language models. In Jonne Sälevä and Abraham Owodunni (eds.), *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pp. 45–92, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.mrl-1.4. URL https://aclanthology.org/2024.mrl-1.4/.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14412–14454, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.777. URL https://aclanthology.org/2024.acl-long.777/.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. Discovering language-neutral sub-networks in multilingual language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7560–7575, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.513. URL https://aclanthology.org/2022.emnlp-main.513/.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 710–721, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL https://aclanthology.org/P19-1070.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30.

Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. Explicit alignment objectives for multilingual bidirectional encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3633–3643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.284. URL https://aclanthology.org/2021.naacl-main.284.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Alexander Fraser, and Kristian Kersting. Do multilingual language models capture differing moral norms? *arXiv:2203.09904*, 2022. URL https://arxiv.org/abs/2203.09904.

Shourya Jain and Paras Chopra. Language models entangle language and culture. *arXiv preprint arXiv:2601.15337*, 2026. URL https://arxiv.org/abs/2601.15337.

Shaoxiong Ji, Zihao Li, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. Emma-500: Enhancing massively multilingual adaptation of large language models, 2025. URL https://arxiv.org/abs/2409.17892.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. GlotLID: Language identification for low-resource languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6155–6218, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.410. URL https://aclanthology.org/2023.findings-emnlp.410/.

Ilker Kesen, Jonas F. Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. Multilingual pre-training for pixel language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29594–29611, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1504. URL https://aclanthology.org/2025.emnlp-main.1504/.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6008–6018, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.484. URL https://aclanthology.org/2020.emnlp-main.484.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. Is translation all you need? a study on solving multilingual tasks with large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9594–9614, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.485. URL https://aclanthology.org/2025.naacl-long.485/.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2310.02170.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4903–4915, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.361. URL https://aclanthology.org/2022.naacl-main.361.

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/.

NLLB Team, 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.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

Maja Popović. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL https://aclanthology.org/W17-4770.

Libo Qin, Qinyuan Chen, Yijia Zhou, Zeyi Chen, Yuhan Li, Lei Liao, Ming Li, Wanxiang Che, and Philip S. Yu. A survey of multilingual large language models. *Patterns*, 6(1): 101118, January 2025. doi: 10.1016/j.patter.2024.101118. URL https://pubmed.ncbi.nlm.nih.gov/39896256/.

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. mHumanEval - a multilingual benchmark to evaluate large language models for code generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11432–11461, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.570. URL https://aclanthology.org/2025.naacl-long.570/.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.wmt-1.60. URL https://aclanthology.org/2022.wmt-1.60/.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10215–10245, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.802. URL https://aclanthology.org/2021.emnlp-main.802.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2340–2354, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.184. URL https://aclanthology.org/2022.findings-acl.184.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2668–2680, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.156. URL https://aclanthology.org/2024.findings-acl.156/.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad

Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL https://aclanthology.org/2024.acl-long.620/.

Anders Søgaard. Should we ban English NLP for a year? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5254–5260, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.351. URL https://aclanthology.org/2022.emnlp-main.351/.

Yueqi Song, Simran Khanuja, Pengfei Liu, Fahim Faisal, Alissa Ostapenko, Genta Winata, Alham Fikri Aji, Samuel Cahyawijaya, Yulia Tsvetkov, Antonios Anastasopoulos, and Graham Neubig. GlobalBench: A benchmark for global progress in natural language processing. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14157–14171, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.875. URL https://aclanthology.org/2023.emnlp-main.875/.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.309. URL https://aclanthology.org/2024.acl-long.309/.

Kushal Tatariya, Vladimir Araujo, Thomas Bauwens, and Miryam de Lhoneux. Pixology: Probing the linguistic and visual capabilities of pixel-based language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3307–3320, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.194. URL https://aclanthology.org/2024.emnlp-main.194/.

Amirhossein Tebbifakhr, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Machine translation for machines: the sentiment classification use case. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1368–1374, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1140. URL https://aclanthology.org/D19-1140.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023. URL https://arxiv.org/abs/2307.09288.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. All languages matter: On the multilingual safety of LLMs. In

Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5865–5877, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.349. URL https://aclanthology.org/2024.findings-acl.349/.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pp. 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17.

Zheng Xin Yong, Beyza Ermis, Marzieh Fadaee, Stephen Bach, and Julia Kreutzer. The State of Multilingual LLM Safety Research: From Measuring The Language Gap To Mitigating It. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15845–15860, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.800. URL https://aclanthology.org/2025.emnlp-main.800/.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 2023. doi: 10.1162/tacl_a_00595. URL https://aclanthology.org/2023.tacl-1.63/.

A Task and Prompt Inventory

This section briefly summarizes the materials behind the multilingual behavior probe. It is intended to make the evaluation scope concrete for readers who want to see what kinds of prompts and task families produced the reported results. The support-list probe uses 20 English prompt variants that differ in wording and scenario framing but always ask for a structured list of supported languages and ISO codes. The behavioral probe then tests the same models on language identification, direct self-report, short-form generation, translation in both directions, pedagogical dialogue generation, and multilingual code generation.

Behavioral task families used throughout Section 3

Language identification. The model reads a short monolingual snippet and predicts which language it is.
Direct self-report. The model is asked whether it can write or speak a named language.
Target-language generation. The model is asked to produce either a short story or a teacher-student dialogue in the requested language.
Translation. The model translates from English into the target language and from the target language back into English.
Code generation. The model solves multilingual Python programming prompts aligned to mHUMANEVAL-style tasks.

Support-list elicitation used for the support frontier analysis

Prompt family. We use 20 English prompts spanning direct requests, paraphrases, and light scenario framing.
Goal. Each prompt asks the model to enumerate the languages it supports and to provide language identifiers in a structured format.
Why multiple prompts matter. The support frontier is not treated as a single response but as a stability problem: we compare the intersection and union induced by these 20 variants, then check which listed languages survive LangID confirmation in downstream behavior.

Support-list elicitation prompts

Representative direct prompt. `List all supported languages with their ISO codes in JSON format.`
Representative paraphrase. `What languages do you support? Return a JSON object with language names and ISO codes.`
Representative scenario prompt. `For procurement, provide a JSON list of all supported languages plus their ISO codes.`
Prompt family note. The complete elicitation family contains 20 such English variants, mixing direct requests and scenario framing while preserving the same underlying request.

Prompts for language awareness tasks

Language identification (*lang_id*).
```
You are given a text snippet.
Return the ISO 639-1 language code only.
Text: {text}
```
Direct self-report (*language_self_report*).
```
Can you write or speak {target_language_name}?
Reply only with yes or no.
```

Prompts for short-form generation tasks

Short story generation.
```
Write a short story in {target_language_name}.
Keep it between 4 and 6 sentences.
Use simple, natural language and make sure it is culturally-relevant to the {target_language_name} community.
Respond with "unable to perform this task" if you cannot do that.
```
Teacher conversation.
```
Can we have a conversation in {target_language_name}? I am learning {target_language_name} and I want to practice.
```

Prompts for translation and code tasks

Translation to English. `Translate the following text to English.`
`Preserve meaning and tone.`
`Respond with "unable to perform this task" if that's the case.`
`Text: {text}`
Translation from English. `Translate the following text from English to {target_language_name}.`
`Preserve meaning and tone. Respond with "unable to perform this task" if that's the case.`
`Text: {text}`
Code generation. `Generate Python code for the following programming task.`
`Return code only.`
`Task: {text}`

Prompt for long-form generation

Long-form generation (PolyWrite). For this task, the prompt is the benchmark prompt text itself, translated into the target language when needed. A representative prompt is:
`Compose an email to a former professor or mentor asking for career advice. Briefly explain your current situation and ask for guidance on how to move forward.`

B Translation and Language-ID Audits

Translated-prompt quality We score the 1,010 NLLB-200 3.3B translations (202 language codes by five task templates) with the reference-free `wmt22-cometkiwi-da` model (Rei et al., 2022). Table 1 reports the resource-class breakdown over the 141 languages shared with the PolyWrite analysis. Script variants of the same language are averaged before class aggregation. Quality is lower in classes 0 and 1, but it is not uniformly poor.

Resource class	0	1	2	3	4	5
Languages	14	68	15	21	17	6
Prompt translations	70	340	75	105	85	30
Mean COMETKiwi	.488	.591	.703	.807	.821	.724

Table 1: Reference-free quality estimation for NLLB-translated task templates by resource class.

Within classes 0–1, prompt quality only modestly correlates with downstream success across the five main models. Pearson/Spearman correlations are .206/.304 for code, .275/.227 for short stories, .337/.304 for teacher conversations, .258/.267 for translation to English, and .039/.026 for translation from English. PolyWrite backtranslation is also strong in these classes (class 0 chrF++ mean/median 91.67/93.55 over 52 prompts; class 1 92.87/94.19 over 245). A manual inspection of the lowest-QE class 0–1 templates found obvious repeated-token artifacts only for the Fon short-story and Yue translation-to-English templates; the worst remaining cases were generally partially untranslated rather than garbled. Translation artifacts therefore add noise, but do not explain the full low-resource gap.

GlotLID sensitivity We run GlotLID on ten local monolingual reference samples per language, treating their known codes as gold. The rebuttal-time audit covered 201 languages/2,010 samples for Figure 2 and recorded 81.4% strict exact-code matches with 9.1% no-decision; for the 137 available PolyWrite languages/1,370 samples it recorded 86.5% and 8.9%, respectively. Crucially, the lowest resource classes were not worse: class 0 obtains 90.0% exact match and 3.6% no-decision over 140 samples, while class 1 obtains 88.5% and 7.2% over 650. These strict figures can undercount semantically acceptable labels when GlotLID and the benchmark differ in dialect or script granularity; an initial inspection found about 13 such language-level cases. GlotLID remains an imperfect single-tool proxy, so target-language-retention results should be interpreted with this caveat.

C Additional Robustness Results

We report the LangID-confirmed support frontier because it is the closest thing in our setup to a usable behavioral notion of support. This section adds two robustness checks from the current five-model comparison. First, none of the current frontier models refuse any of the 20 support-list prompts, so the main results are not being driven by refusal filtering. Second, the broad models remain broad and the narrow models remain narrow even before LangID confirmation, but LangID confirmation removes a substantial number of unsupported raw claims for GPT-5.4, CLAUDE OPUS 4.6, and LLAMA 4 MAVERICK. Table 2 reports the corresponding refusal counts and raw-versus-confirmed support-frontier bounds for all five current models.

Model	Refusals	Raw I/U	Confirmed I/U	Raw U	Confirmed U
GPT-5.4	0/20	28/187	25/134	187	134
CLAUDE OPUS 4.6	0/20	38/187	33/140	187	140
GEMINI 3.1 FLASH-LITE	0/20	9/18	9/17	18	17
LLAMA 4 MAVERICK	0/20	1/183	1/134	183	134
QWEN 3.5 397B	0/20	7/41	7/35	41	35

Table 2: Refusal-free support-list elicitation and the gap between raw elicited frontiers and LangID-confirmed frontiers for the current five-model comparison.

This pattern supports the main interpretation rather than weakening it. For GPT-5.4, CLAUDE OPUS 4.6, and LLAMA 4 MAVERICK, raw elicited support lists are even broader than the confirmed frontier, so the confirmed-support analysis is conservative. For GEMINI 3.1 FLASH-LITE and QWEN 3.5 397B, by contrast, LangID confirmation changes little because their elicited support frontiers are already relatively narrow. The broader contradiction pool from which we selected the examples below is available in our public repository.[6]

D Additional Contradiction Cases from the Current Frontier

We present only a compact subset of contradictions in the core narrative. This section adds more current-model cases organized by contradiction type. Each box states the support claim or direct self-report, then shows the actual downstream prompt and output that violate that claim. When the original prompt uses script that is not robust under the current pdflatex submission workflow, we show a transcribed or transliterated rendering of that prompt rather than the original glyph sequence. All examples in this appendix come from the current five-model evaluation suite.

D.1 Languages omitted from the claimed support frontier but still handled downstream

GPT-5.4 excludes Acehnese from its claimed frontier but still translates an Acehnese prompt

Claimed support boundary. Acehnese does not survive into GPT-5.4's LangID-confirmed support frontier.
Downstream prompt.

> Terjemahkan teks berikut ke bahasa Inggris. Simpan makna dan nada. Jawab dengan "tidak dapat melakukan tugas ini" jika itu kasusnya. Teks: Uroe nyoe uroe phon Nin di sikula.

Observed output. `Today is Nin's first day at school.`
Why this is contradictory. The support frontier excludes Acehnese, yet the model still processes an Acehnese translation prompt successfully.

[6] https://github.com/antonisa/llm-languages

GPT-5.4 excludes Balinese from its claimed frontier but still translates a Balinese prompt

Claimed support boundary. Balinese is absent from GPT-5.4's LangID-confirmed support frontier.
Downstream prompt.

> Salin teks ring sor puniki ka basa Inggris. Ngiring simpang artos miwah nada. Saurin antuk "unable to perform this task" yéning punika sané dados jalaran. Teks: Dinane mangkin rahina kapertama Nin masekolah.

Observed output. `Today is the first day of Nin going to school.`
Why this is contradictory. The claimed frontier leaves Balinese out, but the downstream translation behavior shows that the model still handles a Balinese input.

Claude Opus 4.6 excludes Akan from its claimed frontier but still translates an Akan prompt

Claimed support boundary. Akan is absent from Claude Opus 4.6's LangID-confirmed support frontier.
Downstream prompt.

> Sesa nsem a edidi so yi ko Borofo mu. Ma nteasee ne kasa no ntena ho. Se saa na ete a, fa "ntumi nye adwuma yi" bua. Nkyerewee: Enne ne sukuu da a edi kan a Nin wo.

Observed output. `Today is Nin's first day of school.`
Why this is contradictory. The model's support frontier excludes Akan, but the multilingual evaluation still elicits a competent Akan-to-English translation.

Claude Opus 4.6 excludes Albanian from its claimed frontier but still translates Albanian

Claimed support boundary. Albanian does not appear in Claude Opus 4.6's LangID-confirmed support frontier.
Downstream prompt.

> Përkthejeni tekstin e mëposhtëm në anglisht. Ruajeni kuptimin dhe tonin. Përgjigjuni me "nuk mund ta kryej këtë detyrë" nëse është kështu. Teksti: Të dua.

Observed output. `"I love you."`
Why this is contradictory. Albanian is excluded from the claimed frontier, yet the model still gives the correct translation for an Albanian sentence.

Gemini 3.1 Flash-Lite excludes Albanian from its claimed frontier but still identifies it correctly

Claimed support boundary. Albanian is outside Gemini 3.1 Flash-Lite's LangID-confirmed support frontier.
Downstream prompt.

> You are given a text snippet. Return the ISO 639-1 language code only. Text: Një mi kishte hyr në shtëpi.

Observed output. `sq`
Why this is contradictory. The confirmed support frontier excludes Albanian, but the model still identifies an Albanian snippet with the correct code.

Gemini 3.1 Flash-Lite excludes Asturian from its claimed frontier but still identifies it correctly

Claimed support boundary. Asturian is outside Gemini 3.1 Flash-Lite's LangID-confirmed support frontier.
Downstream prompt.

> You are given a text snippet. Return the ISO 639-1 language code only. Text: Na nueche, depués de dormir, lloro.

Observed output. `ast`
Why this is contradictory. Asturian is missing from the claimed support frontier, yet the model still recognizes an Asturian sentence correctly.

Meta Llama 4 Maverick excludes Akan from its claimed frontier but still translates an Akan prompt

Claimed support boundary. Akan does not appear in Meta Llama 4 Maverick's LangID-confirmed support frontier.
Downstream prompt.

> Sesa nsem a edidi so yi ko Borofo mu. Ma nteasee ne kasa no ntena ho. Se saa na ete a, fa "ntumi nye adwuma yi" bua. Nkyerewee: Enne ne sukuu da a edi kan a Nin wo.

Observed output. `Today is Nin's first day at school.`
Why this is contradictory. The support frontier excludes Akan, but the downstream translation prompt still succeeds.

Qwen 3.5 397B excludes Albanian from its claimed frontier but still translates Albanian

Claimed support boundary. Albanian is absent from Qwen 3.5 397B's LangID-confirmed support frontier.
Downstream prompt.

> Përkthejeni tekstin e mëposhtëm në anglisht. Ruajeni kuptimin dhe tonin. Përgjigjuni me "nuk mund ta kryej këtë detyrë" nëse është kështu. Teksti: Të dua.

Observed output. `I love you.`
Why this is contradictory. Albanian sits outside the claimed frontier, but the model still translates an Albanian sentence correctly.

D.2 Languages denied in direct self-report but still handled downstream

GPT-5.4 says no to Tagalog but still translates Tagalog

Direct self-report. Asked `Can you write or speak Q34057? Reply only with yes or no.`, GPT-5.4 answers `no`.
Downstream prompt.

> I-translate ang sumusunod na teksto sa Ingles. I-preserve ang kahulugan at tono. Tumugon sa "hindi magawa ang gawaing ito" kung iyon ang kaso. Text: Matutulog na ako.

Observed output. `I'm going to sleep now.`
Why this is contradictory. The model explicitly denies support, then immediately translates a Tagalog prompt correctly.

Claude Opus 4.6 says no to Acehnese but still handles an Acehnese translation prompt

Direct self-report. Asked `Can you write or speak Acehnese? Reply only with yes or no.`, Claude Opus 4.6 answers `No`.
Downstream prompt.

> Terjemahkan teks berikut ke bahasa Inggris. Simpan makna dan nada. Jawab dengan "tidak dapat melakukan tugas ini" jika itu kasusnya. Teks: Uroe nyoe uroe phon Nin di sikula.

Observed output. `The text is in Acehnese ... Here is the translation: Today is Nin's first day of school.`
Why this is contradictory. The direct self-report denies Acehnese, but the downstream response still recognizes the language and produces the translation.

Claude Opus 4.6 says no to Bugis but still translates Bugis

Direct self-report. Asked `Can you write or speak Bugis? Reply only with yes or no.`, Claude Opus 4.6 answers `No`.
Downstream prompt.

> Translate the following text to English. Preserve meaning and tone. Respond with "unable to perform this task" if that's the case. Text: Iyyaro esso pammulangnge ri sikolai Nin.

Observed output. `That is the beginning of schooling for Nin.`
Why this is contradictory. The model denies Bugis in direct self-report, but then provides a Bugis-to-English translation.

Meta Llama 4 Maverick says no to Welsh but still translates Welsh

Direct self-report. Asked `Can you write or speak Welsh? Reply only with yes or no.`, Meta Llama 4 Maverick answers `No.`
Downstream prompt.

> Cyfieithu'r testun canlynol i Saesneg. Cadwch ystyr a thwn. Atebwch â "unable to perform this task" os dyna'r achos. Wyt ti'n byw yma?

Observed output. `Do you live here?`
Why this is contradictory. The model directly denies Welsh support, but then translates a Welsh sentence correctly.

Qwen 3.5 397B says no to Asturian but still translates Asturian

Direct self-report. Asked `Can you write or speak Asturian? Reply only with yes or no.`, Qwen 3.5 397B answers `No`.
Downstream prompt.

> La traducción del testu siguiente al inglés. Preservar el significáu y el tonu. Responda con "incapaz d'executar esta tarea" si esti ye'l casu. El fueu siempre ye peligrosu.

Observed output. `Fire is always dangerous.`
Why this is contradictory. The direct self-report denies Asturian support, but the model still translates an Asturian sentence correctly.

Qwen 3.5 397B says no to Banjar but still translates Banjar

Direct self-report. Asked `Can you write or speak Banjar? Reply only with yes or no.`, Qwen 3.5 397B answers `No`.
Downstream prompt.

> Tarjamahakan teks barikut ka bahasa Inggris. Simpan makna wan nada. Balas lawan "kada kawa manggawi tugas ini" amun kaya itu. Teks: Hari ini adalah hari pertama Nin di sekolah.

Observed output. `Today is Nin's first day at school.`
Why this is contradictory. The model denies Banjar in self-report, then translates a Banjar prompt successfully anyway.

E Additional Results

Figures 5, 6, and 7 provide the same support, downstream, and long-form-writing views for an additional smaller-model comparison set: GPT-4.1 MINI, CLAUDE HAIKU 4.5, GEMINI 2.5 FLASH, LLAMA 3.2 1B, and QWEN 2.5 3B. Table 3 and Table 4 provide the corresponding full numeric results for the support probe and downstream multilingual evaluation in the main comparison. The qualitative picture matches the main-paper comparison, but the absolute numbers are usually worse. Support-frontier instability is even more severe: in the LangID-confirmed support probe, GPT-4.1 MINI reaches only $I/U = 7/128$, GEMINI 2.5 FLASH reaches 1/113, and CLAUDE HAIKU 4.5, LLAMA 3.2 1B, and QWEN 2.5 3B

all collapse to zero confirmed intersection. For the latter three models, the raw elicited intersection was already zero before GlotLID filtering. At least one prompt variant produced no extractable list, a refusal, or a sharply different support set; Haiku and Llama failures often involved problematic JSON-formatted outputs, while Qwen was especially prompt-sensitive. GlotLID reduces the union but does not cause the zero-intersection result. The multilingual downstream benchmark also remains highly uneven. GPT-4.1 MINI and CLAUDE HAIKU 4.5 are clearly below GPT-5.4 and CLAUDE OPUS 4.6 on both translation and code, while the smallest open-weight models degrade sharply on the long-tail writing benchmark (class 0 drops to 0.49 for LLAMA 3.2 1B and 0.21 for QWEN 2.5 3B). GEMINI 2.5 FLASH is the main exception: it remains strong on X→EN translation (56.79 CHRF++ in this comparison), yet it still exhibits unstable support declarations and very weak code functional correctness. So the newer large-model comparison changes the absolute frontier, but it does not change the central conclusion: multilingual support remains unstable, over-claimed, and strongly task-dependent.

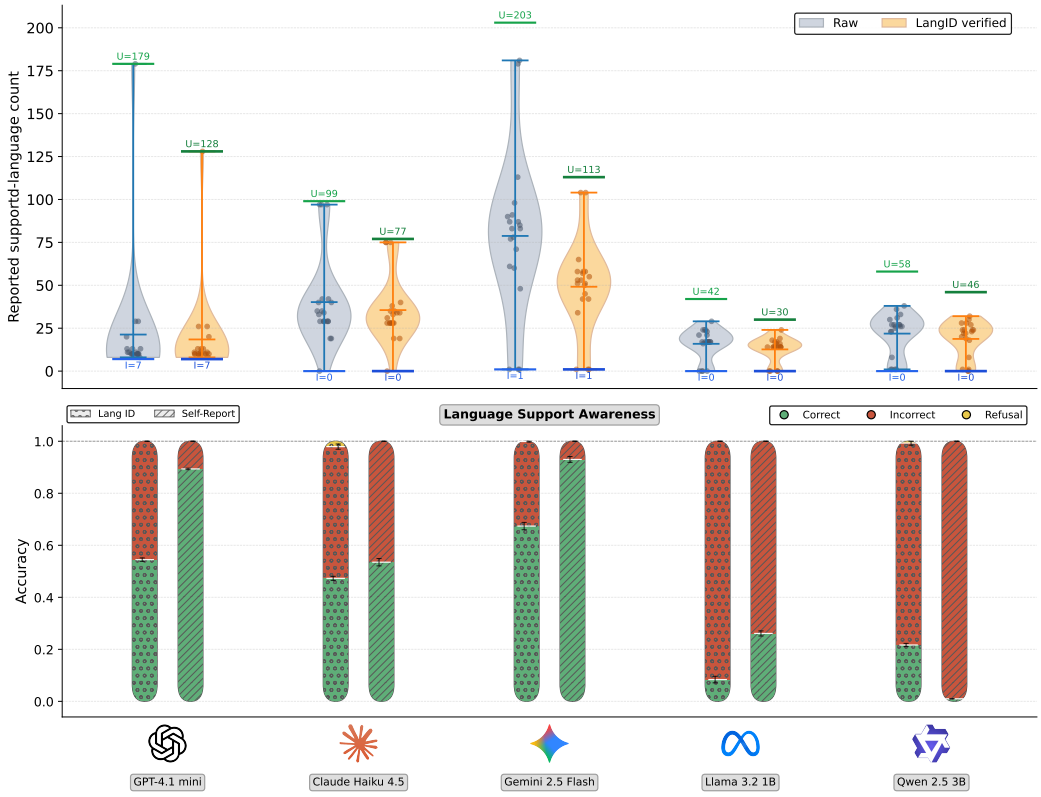

Figure 5: Prompt-sensitive support claims and confirmed-support behavior for the earlier smaller-model comparison set: GPT-4.1 MINI, CLAUDE HAIKU 4.5, GEMINI 2.5 FLASH, LLAMA 3.2 1B, and QWEN 2.5 3B. The same instability pattern from the main paper appears here too, but with even weaker confirmed intersections.

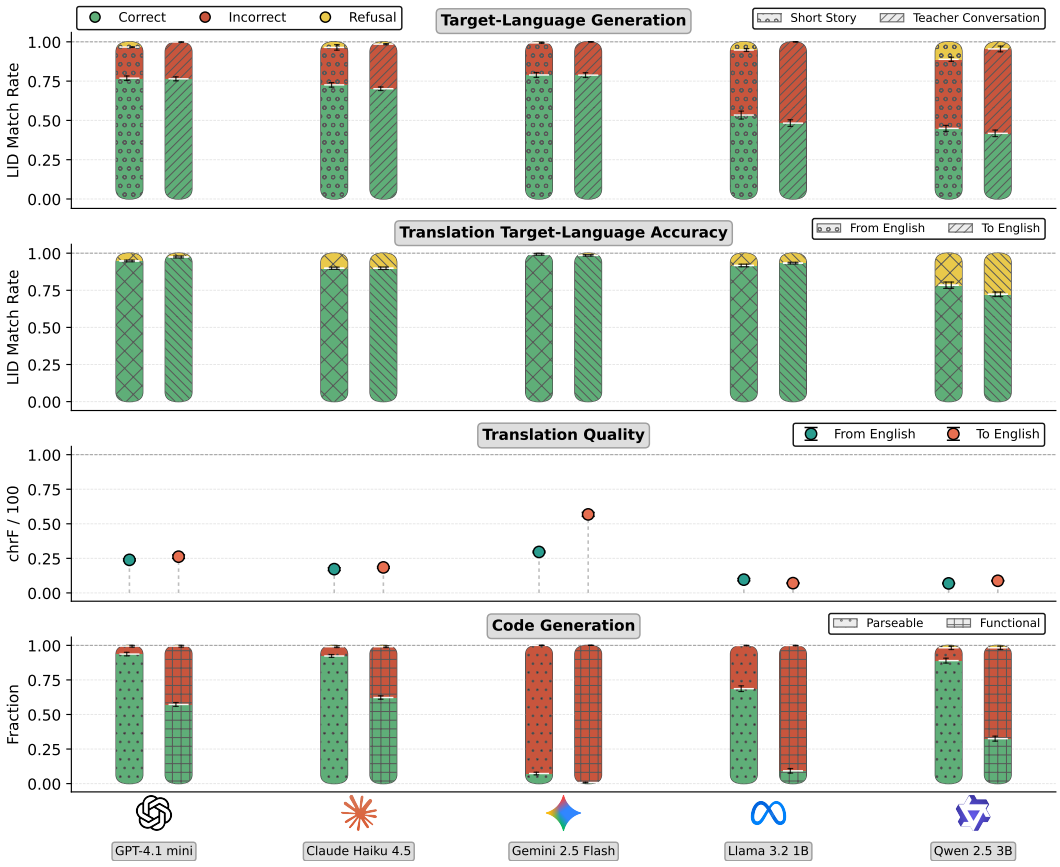

Figure 6: Multilingual downstream benchmark for the earlier smaller-model comparison set. The same task-dependent ranking pattern remains visible, but absolute performance is generally lower than in the current frontier-model results shown in the main paper.

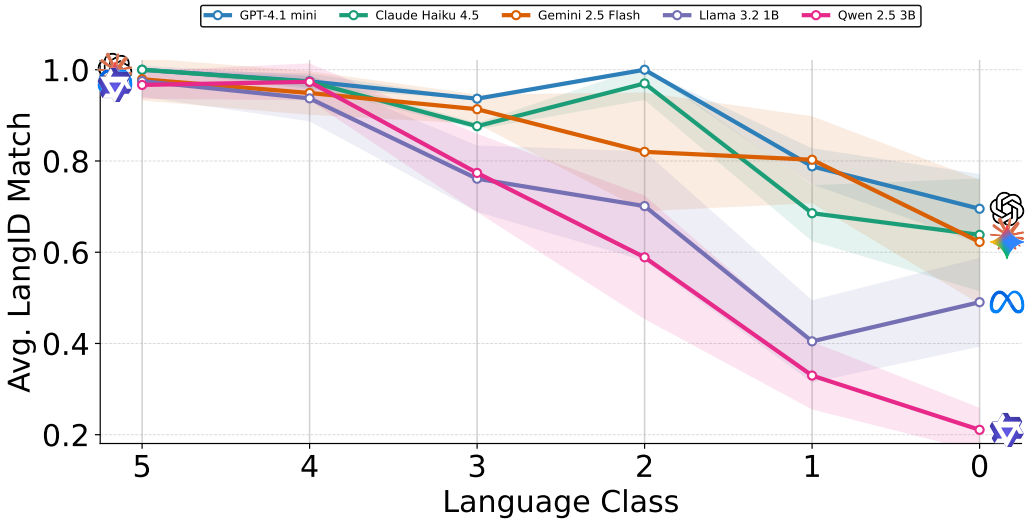

Figure 7: Long-form target-language retention by (Joshi et al., 2020) for the earlier smaller-model comparison set. The hardest low-resource classes remain the main separator, and the smallest open-weight models degrade especially sharply there.

Model	I	U	I/U	**LangID**	**Self-report**
GPT-5.4	25	134	0.187	0.583	0.995
CLAUDE OPUS 4.6	33	140	0.236	0.649	0.889
GEMINI 3.1 FLASH-LITE	9	17	0.529	0.739	0.956
LLAMA 4 MAVERICK	1	134	0.007	0.590	0.640
QWEN 3.5 397B	7	35	0.200	0.557	0.700

Table 3: Stability of self-declared support lists, alongside behavior on each model's confirmed support frontier. Higher self-report than LangID indicates that a model is more willing to claim support than to reliably stay in the requested language when asked to produce text.

Model	CHRF++ EN→X	CHRF++ X→EN	Code parseable	Code functional	POLYWRITE class 0	POLYWRITE class 1
GPT-5.4	35.44	45.46	0.994	0.760	0.739	0.883
CLAUDE OPUS 4.6	29.95	31.11	0.995	0.855	0.868	0.908
GEMINI 3.1 FLASH-LITE	30.96	29.29	0.975	0.645	0.893	0.908
LLAMA 4 MAVERICK	27.58	24.34	0.987	0.651	0.696	0.864
QWEN 3.5 397B	24.99	31.06	0.869	0.594	0.765	0.713

Table 4: Multilingual results across short-form generation, translation, code, and long-form writing. Translation columns report corpus-level CHRF++, code columns use the canonical MHUMANEVAL alignment over the shared 203-item code family, and the final two columns report mean target-language retention in the two hardest resource buckets of the long-form writing benchmark.