MAPS: A Multilingual Benchmark for Global Agent Performance and Security

Omer Hofman[1], Oren Rachmil[1], Shamik Bose[1], Vikas Pahuja[1], Jonathan Brokman[1],
Toshiya Shimizu[2], Trisha Starostina[3], Kelly Marchisio[3],
Seraphina Goldfarb-Tarrant[3], Roman Vainshtein[1]

[1]Fujitsu Research of Europe [2]Fujitsu Limited [3]Cohere

Abstract

Agentic AI systems, which build on Large Language Models (LLMs) and interact with tools and memory, have rapidly advanced in capability and scope. Yet, since LLMs have been shown to struggle in multilingual settings, typically resulting in lower performance and reduced safety, agentic systems risk inheriting these limitations. This raises concerns about the global accessibility of such systems, as users interacting in languages other than English may encounter unreliable or security-critical agent behavior. Despite growing interest in evaluating agentic AI, existing benchmarks focus exclusively on English, leaving multilingual settings unexplored. To address this gap, we propose MAPS, a multilingual benchmark suite designed to evaluate agentic AI systems across diverse languages and tasks. MAPS builds on four widely used agentic benchmarks — GAIA (real-world tasks), SWE-bench (code generation), MATH (mathematical reasoning), and the Agent Security Benchmark (security). We translate each dataset into ten diverse languages, resulting in 805 unique tasks and 8,855 total language-specific instances. Our benchmark suite enables a systematic analysis of how multilingual contexts affect agent performance and robustness. Empirically, we observe consistent degradation in both performance and security when transitioning from English to other languages, with severity varying by task and correlating with the amount of translated input. Building on these findings, we provide actionable recommendations to guide agentic AI systems development and assessment under multilingual settings. This work establishes a standardized evaluation framework, encouraging future research towards equitable, reliable, and globally accessible agentic AI. MAPS benchmark suite is publicly available. [1]

1 Introduction

LLM-based agentic AI systems combine multi-step reasoning with external tools and memory to solve open-ended tasks such as code generation, web navigation, planning, and transactional services like booking and ordering [Acharya et al., 2025]. By doing so, they extend to complex, real-world problems beyond standard LLM benchmarks. Since such real-world applications serve speakers of diverse languages, maintaining consistent reliability in every language becomes critical. However, since agentic behavior is grounded in LLMs, which often perform inconsistently across languages [Deng et al., 2023, Wang et al., 2023], agents may inherit these multilingual limitations as well, affecting their functionality and trustworthiness. This presents a barrier to equitable access, as non-English users may face degraded responses, incorrect tool actions, or unsafe behaviors—failures that can lead to actual harm in the real world, including erroneous transactions, data corruption, and other security vulnerabilities [Zhang et al., 2024].

[1] `https://huggingface.co/datasets/Fujitsu-FRE/Multilingual-Agentic-AI-Benchmark`

To assess emerging agentic systems, various benchmarks have been proposed to evaluate agent performance across a range of tasks [Mialon et al., 2023, Jimenez et al., 2023, Chang et al., 2024, Xu et al., 2024]. However, these benchmarks remain English-only. In contrast to multilingual LLM benchmarks [Dang et al., 2024, Shi et al., 2022, Goyal et al., 2022], no equivalent exists for agentic AI tasks—creating a blind spot in our understanding of cross-language performance, safety, and security.

In this paper, we address this gap. We hypothesize that multilingual settings will reveal performance and security gaps in agentic systems that are not captured by the existing, English-only benchmarks. To investigate this, we introduce MAPS, a **M**ultilingual **A**gentic AI Benchmark Suite for **P**erformance and **S**ecurity. MAPS is based on four established agentic benchmarks: GAIA (real-world tasks) [Mialon et al., 2023], SWE-bench (code generation) [Jimenez et al., 2023], MATH (mathematical reasoning) [Hendrycks et al., 2021], and the Agent Security Benchmark (security) [Zhang et al., 2024]. These benchmarks are extended to ten typologically diverse languages beyond English[2], by employing a hybrid machine and LLM-based translation approach Ki and Carpuat [2024] with extended verification and enhancements. In total, MAPS includes 805 unique tasks, each available in 11 language versions—including the original English and 10 translated variants—for a total of 8,855 instances. An overview of the benchmark structure is shown in Figure 1.

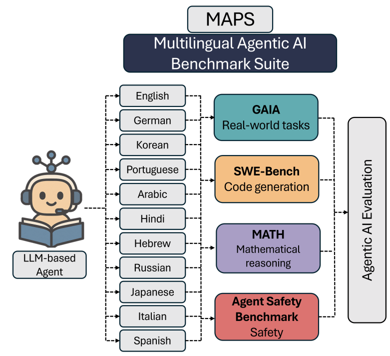

Figure 1: MAPS benchmark suite evaluates LLM-based agents across 11 languages and 4 agentic benchmarks covering performance and security.

To demonstrate the use of MAPS and test our hypothesis, we selected a leading open-source agent associated with each of the four original benchmarks and applied it to the corresponding multilingual extension. We observed notable declines in both task performance and security when moving from English to other languages, with the severity of these drops varying by task type and correlating with the proportion of non-English input, suggesting that multilingual performance interventions should be targeted based on input composition and task sensitivity. Beyond overall degradation, our findings reveal that multilingual inputs can amplify agentic vulnerabilities in safety-critical tasks, highlighting the need for multilingual risk assessment. These results empirically support our hypothesis and demonstrate the utility of MAPS as a tool for systematic, multilingual evaluation of agentic AI systems.

The primary contributions of this paper are threefold:

- To the best of our knowledge, we introduce the first multilingual benchmark suite for agentic AI, extending four widely used benchmarks into ten typologically diverse languages for systematic performance and security assessment.
- The efficacy and quality of the proposed benchmark are demonstrated through a large-scale evaluation of leading agents as well as human expert verification.
- We present the first quantifiable analysis and evidence that multilingual settings reveal critical performance, safety, and security gaps in agentic systems, along with actionable recommendations for improving their development.

2 Background and Related Work

2.1 Agentic AI Benchmarks

With the rapid advancement of LLM-based agents, a diverse suite of benchmarks has been developed to assess their autonomy, tool use, planning, and memory integration [Yao et al., 2024, Xu et al., 2024, Yehudai et al., 2025]. We organized these suites along three primary dimensions.

[2] MAPS offers evaluation in the following ten languages: German, Spanish, Portuguese (Brazil), Japanese, Russian, Italian, Arabic, Hebrew, Korean, and Hindi.

Evaluation objective: performance-oriented benchmarks measure task completion, multi-step reasoning, and correct tool invocation (e.g., AgentBoard [Chang et al., 2024]), whereas security-focused suites probe robustness to adversarial inputs, jailbreaks, and unsafe behaviors (e.g., AgentHarm [Andriushchenko et al., 2024]). **Agentic task scope**: full-agentic evaluations present only problem statements and expected outcomes, requiring end-to-end planning and execution (e.g., GAIA [Mialon et al., 2023]), while semi-agentic frameworks supply scaffolding, such as code templates or mock APIs, to isolate the LLM's reasoning and tool-selection core (e.g., AppWorld [Trivedi et al., 2024]. **Design and evaluation characteristics**: most benchmarks span a limited set of domains (three to five use cases), typically including real-world information retrieval and navigation (e.g., AssistantBench [Yoran et al., 2024]), code generation (e.g., SWE-Bench [Jimenez et al., 2023]), reasoning and planning (e.g., MATH [Hendrycks et al., 2021], Travel Planner [Xie et al., 2024]), and security scenarios (e.g., Agent Security Benchmark [Zhang et al., 2024]). They use fixed task counts and predefined difficulty tiers, and to enable reliable, objective measurement despite agents' open-ended capabilities, they often restrict tasks to closed-form problems with definitive ground truth, allowing clear determination of success or failure [Jimenez et al., 2023, Mialon et al., 2023]. A detailed comparison of benchmark design choices, task types, and evaluation properties is provided in the supplementary materials.

While multilingual LLM benchmarks such as XTREME [Hu et al., 2020], FLORES [Goyal et al., 2022], and SIB-200 [Adelani et al., 2023] have enabled broad cross-lingual evaluation, they do not assess interactive decision-making, tool use, or task execution, which are core elements of agentic systems. As a result, existing multilingual benchmarks fall short of capturing the complexities and vulnerabilities that arise when agents operate in non-English settings. This leaves non-English users exposed to agentic failures in their native languages and underscores the need for fully agentic benchmarks that include performance and security evaluation, high data fidelity, and comprehensive multilingual assessment - gaps that our benchmark is specifically designed to address.

2.2 Multilingual Benchmarks and Multilingual Limitations of General-Purpose LLMs

Recent studies show that pre-trained LLMs often struggle with non-English input, especially in languages with limited training resources or those typologically distant from English. Multilingual benchmarks such as XTREME [Hu et al., 2020] and XGLUE[Liang et al., 2020] report consistent accuracy drops when moving from English to languages such as Swahili or Nepali. These gaps reflect an imbalance in pretraining corpora, where English accounts for over 90% of the data, as well as challenges in tokenizing morphologically rich languages and the scarcity of fine-tuning data in many languages [Jha, 2024]. Notably, even large models (e.g., GPT-4, Llama 405B) face a "cross-lingual knowledge barrier" on MMLU [Hendrycks et al., 2020] and on Safety tasks [Grattafiori et al., 2024], showing that scale alone does not resolve multilingual performance deficits [Chua et al., 2024].

Building on these performance gaps, LLMs also face robustness and security challenges in multilingual contexts. Since most alignment and red-teaming efforts have been English-centric, models are more prone to generate toxic or policy-violating outputs when processing non-English prompts [Deng et al., 2023, Wang et al., 2023, Aakanksha et al., 2024]. Furthermore, hallucination rates increase and confidence calibration degrades outside English, causing models to produce fluent, yet unreliable or potentially harmful content in undersupported languages [Xue et al., 2024]. Although security interventions, such as multilingual alignment, have been shown to be effective in reducing harmful output between languages, they often incur a measurable cost in downstream performance or increased latency [Aakanksha et al., 2024].

Given that agentic AI systems are based directly on these LLMs, we hypothesize that they inherit the same language-dependent performance and security limitations. As these agents carry out real-world tasks such as executing code, querying external tools, or navigating web environments, any inherited shortcomings can lead to severe consequences. Yet, to our knowledge, no systematic evaluation has probed how these vulnerabilities manifest within agentic systems. To address this gap, we introduce MAPS, our multilingual agentic benchmark suite in Section 3.

3 MAPS: Multilingual Agentic AI Benchmark Suite

To support multilingual evaluation of agentic systems, we construct a benchmark suite by extending established English-language datasets into multiple languages. This process requires careful dataset

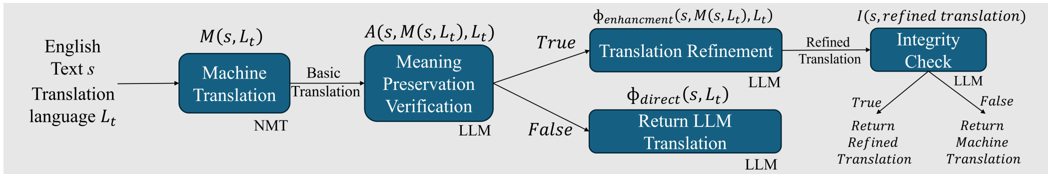

Figure 2: **Overview of our multi-stage translation pipeline for agentic benchmark construction.** We start with machine translation for structural alignment, followed by LLM-based verification and enhancement. This approach is adapted from Ki and Carpuat [2024] but extended with task-specific prompting and fallback mechanisms tailored to the requirements of agentic AI evaluation.

selection, translation procedures that preserve semantic and structural integrity, and mechanisms for ensuring evaluation consistency. The following subsections detail our methodology for translation, benchmark construction, and dataset composition.

3.1 Translation Pipeline

Reliable multilingual evaluation of agentic AI systems hinges on translating task instructions with both semantic and structural cross-language fidelity. Neural MT excels at preserving format and structure, but struggles in low-resource or specialized domains [Koehn and Knowles, 2017, Aharoni et al., 2019]. Translation via instructed LLMs offers broader high-level capabilities at the cost of occasional hallucinations and semantic drift [Hendy et al., 2023, Yan et al., 2024]. To balance these trade-offs, hybrid pipelines were suggested by Ki and Carpuat [2024], Mohammed [2025], combining format-preserving MT with LLM-based refinement. For MAPS, we extend Ki and Carpuat [2024]: First, Ki and Carpuat [2024] was not designed with our benchmarks in mind, thus significant per-benchmark prompting had to be done. Second, we added automated quality checks, fallbacks, and expert verification to ensure the cross-language fidelity needed for agentic benchmarks (Fig. 2).

Formally, let us express our translation pipeline as a function $T : \mathcal{S} \times \mathcal{L} \to \mathcal{T}$, where $s \in \mathcal{S}$ is a task-instruction instance in source-language (English), $L_t \in \mathcal{L}$ is the target language, and $t \in \mathcal{T}$ is the resulting translated output. The pipeline begins with machine translation (MT) to establish a structural foundation: Denote $M(s, L_t)$, the MT function, implemented as a high-quality, off-the-shelf NMT system. Its output provides a structurally faithful baseline for subsequent steps.

Next, we apply a verification step using an LLM to assess whether the translation adequately preserves the source meaning. This is modeled as a binary function $A(s, M(s, L_t), L_t) \to 0, 1$, where the LLM compares the original and translated texts to detect major semantic errors or omissions.

Based on verification outcomes, the pipeline follows one of two distinct paths. If $(A = 0)$, indicating machine translation failure, the pipeline employs direct LLM translation: Denote $\Phi_{\text{direct}}(s, L_t)$ the output of an LLM prompted to directly translate s to language L_t (without using the MT output). If $(A = 1)$ indicating acceptable machine translation, an LLM enhances the translation while preserving its basic structure: Denote $\Phi_{\text{enhancement}}(s, M(s, L_t), L_t)$ as the output of an LLM, guided to refine and improve the MT output while maintaining structural consistency.

To ensure semantic integrity, we apply a second binary check: $I(s, \Phi_{\text{enhancement}}) \to 0, 1$. This integrity check detects common LLM failure modes, such as hallucinations, omissions, misinterpretations (e.g., answering instead of translating), and semantic drift. If this verification fails, we revert to the original machine translation (which passed the initial verification test).

The added conditional steps form a robust decision framework: If machine translation is rejected, we use a direct LLM translation; if accepted but enhancement fails integrity verification, we fall back to the machine translation; Otherwise, we use the enhanced translation. This structure ensures graceful degradation, favoring conservative outputs when refinement is unreliable. Formally, the final translation is given by:

$$
T(s, L_t) = \begin{cases} \Phi_{\text{direct}}(s, L_t), & \text{if } A(s, M(s, L_t), L_t) = 0 \\ M(s, L_t), & \text{if } A(s, M(s, L_t), L_t) = 1 \text{ and } I(s, \Phi_{\text{enhancement}}(s, M(s, L_t), L_t)) = 0 \\ \Phi_{\text{enhancement}}(s, M(s, L_t), L_t), & \text{otherwise} \end{cases} . \tag{1}
$$

To ensure the reliability of this pipeline across languages and task types, we conducted human verification on a representative subset of translations. Evaluation design and results are detailed in Subsection 3.3, with additional implementation details in the Supplementary Material.

3.2 Dataset Selection and Composition

Dataset Selection. To support robust multilingual evaluation across agentic capabilities, we construct MAPS benchmark suite based on established agentic AI benchmarks. Those were selected based on four criteria: (1) strong adoption and recognition within the research community, including prior use in agentic evaluation; (2) clearly defined, closed-form answers to enable controlled evaluation; (3) sufficient difficulty to challenge current open-source agents without saturating performance; and (4) practical solvability, ensuring that multilingual degradation can be meaningfully measured. Based on these criteria, we selected four datasets spanning real-world reasoning, software engineering, mathematical problem solving, and security assessment.

GAIA. GAIA [Mialon et al., 2023] is a benchmark designed to evaluate agents' performance on real-world assistant tasks. It includes curated questions that require multi-step reasoning and autonomous use of tools such as web browsers, code interpreters, or document analyzers. Each question has a single correct answer, and responses are evaluated by an exact match to a reference output.

SWE-bench. SWE-bench [Jimenez et al., 2023] is a software engineering benchmark constructed from real GitHub issues and associated pull requests across popular Python repositories. Each task presents a bug report and a codebase snapshot, and requires the agent to evaluate whether a proposed patch correctly resolves the issue. We adopt the *verified* subset[3], in which agents are tasked with validating a patch rather than generating one.

MATH. The MATH dataset [Hendrycks et al., 2021] includes high-school level mathematical problems across seven topics, including algebra, geometry, and calculus. Tasks are structured to require symbolic manipulation and multi-step reasoning. Agent responses are evaluated by exact match against a reference solution.

Agent Security Benchmark (ASB). ASB benchmark [Zhang et al., 2024] provides a structured evaluation of agent robustness against adversarial threats, including prompt injections, memory poisoning, and tool misuse. Agents interact with injected prompts or environments, and evaluation is based on whether safety policies are violated, measured by attack success rate and refusal rate.

Data Composition. The metadata below summarizes the multilingual extension, including language coverage, scale, and pre-processing.

Translated Languages. We selected the following ten typologically and geographically diverse languages: German, Spanish, Portuguese (Brazil), Japanese, Russian, Italian, Arabic, Hebrew, Korean, and Hindi. This selection enables the evaluation of agent performance across a wide range of scripts, linguistic structures, and regional user populations.

Dataset Handling. To preserve the integrity and utility of the original datasets, we applied only minimal and targeted modifications. Across all datasets, we appended translations without modifying or removing any original metadata (such as task type, difficulty level, tools available, etc). Domain-specific syntax—such as equations in MATH, code snippets in SWE-bench, and adversarial prompts in ASB—was preserved exactly, maintaining the original task structure and technical fidelity. For MATH and SWE-bench, which were not originally designed for agentic evaluation, we further applied selective filtering to retain only the most challenging tasks based on the task difficulty field. This follows common practice in prior work to align non-agentic datasets with agentic evaluation settings [Wu et al., 2023], by ensuring meaningful evaluation of agent behavior while avoiding trivial cases.

[3] `https://openai.com/index/introducing-swe-bench-verified/`

Data Volume. To balance performance and security evaluation, our benchmark comprises 805 tasks: 405 from performance-oriented datasets (GAIA, SWE-bench, MATH) and 400 from the Agent Security Benchmark. We selected 165 tasks from GAIA (full validation set), 140 high-difficulty tasks from MATH (20 per topic across 7 topics), and 100 hard and medium tasks from SWE-bench. The remaining 400 tasks include all security-relevant prompts from ASB. Each task was translated into 10 target languages, and combined with the original English version, this results in a total of $8,855$ multilingual tasks across 11 languages.

To validate the benchmark's utility and examine multilingual effects, we applied a leading agent to each dataset. Full details and performance results are reported in Section 4.

3.3 Translation Implementation and Verification

Translation Implementation Details. We implemented the hybrid translation pipeline described in Section 3.1 using a combination of commercial and open-source tools. For machine translation, we used the Google Translate NMT API [4], selected for its support across all ten target languages. To preserve task fidelity, structural elements (e.g., equations, variables, code) that MT systems often mistranslate were temporarily masked and restored after translation. For LLM-based refinement and quality control, we used Cohere Command A and GPT-4o, both multilingual models executed with deterministic decoding (temperature set to zero) to ensure output consistency. System prompts were crafted individually for each task to accommodate domain-specific structures (e.g., code snippets, equations, web URLs), ensuring that the models preserved both intent and format. The code is available [5] and representative examples of these prompts are provided in the Supplementary Material.

Human Verification Protocol. To assess translation quality, we manually verified a representative sample of $2,000$ translations, covering 25% of the benchmark, proportionally sampled across datasets and languages. Each item was rated by a bilingual annotator fluent in English and the target language on a $1\breve{}5$ Likert scale across three criteria: *adequacy* (semantic fidelity), *fluency* (grammatical and stylistic naturalness), and *formatting accuracy* (preservation of LaTeX, code, etc.). A fourth metric, *answerability*, measured whether the translation preserved intent well enough for the annotator to confidently answer the question as if it were in English. Annotator instructions are provided in the Supplementary Material. To validate the reliability of the verification process, we embedded "honeypot" samples with intentional errors; annotators reliably flagged these cases, confirming attentiveness and quality control.

Evaluation results confirm high translation quality across the benchmark, with an answerability rate of 94.4%, corresponding to a total error rate of 5.6%. Translations also received average scores of 4.47 for adequacy, 4.60 for fluency, and 4.75 for formatting accuracy (on a 1–5 Likert scale), supporting the benchmark's preservation of semantic fidelity, linguistic naturalness, and structural integrity. Full per-language results and analysis are included in the Supplementary Material. To support high-precision use cases, we also release a "verified"[6] subset of the benchmark, consisting of 190 translations per language that passed human review across all four datasets.

4 Experiments

We now demonstrate MAPS utility through multilingual evaluation of leading open-source agents.

4.1 Experimental Settings

Agent Assignment per Dataset. To demonstrate the utility of our benchmark, we evaluate open-source agents on each dataset and assess their performance and robustness under multilingual settings. While a unified agent would offer a more broad coverage and controlled evaluation, current systems lack the generalization needed to perform well across diverse tasks [Gioacchini et al., 2024, Chang et al., 2024]. To isolate the effect of language variation, we retain each agent's original configuration, including tools, prompts, and system settings, without any modification.

[4] `https://cloud.google.com/translate/docs/advanced/translating-text-v3#translating_input_strings`

[5] `https://github.com/omerhof-fujitsu/hybrid_translation_demo`

[6] `https://huggingface.co/datasets/Fujitsu-FRE/MAPS_Verified`

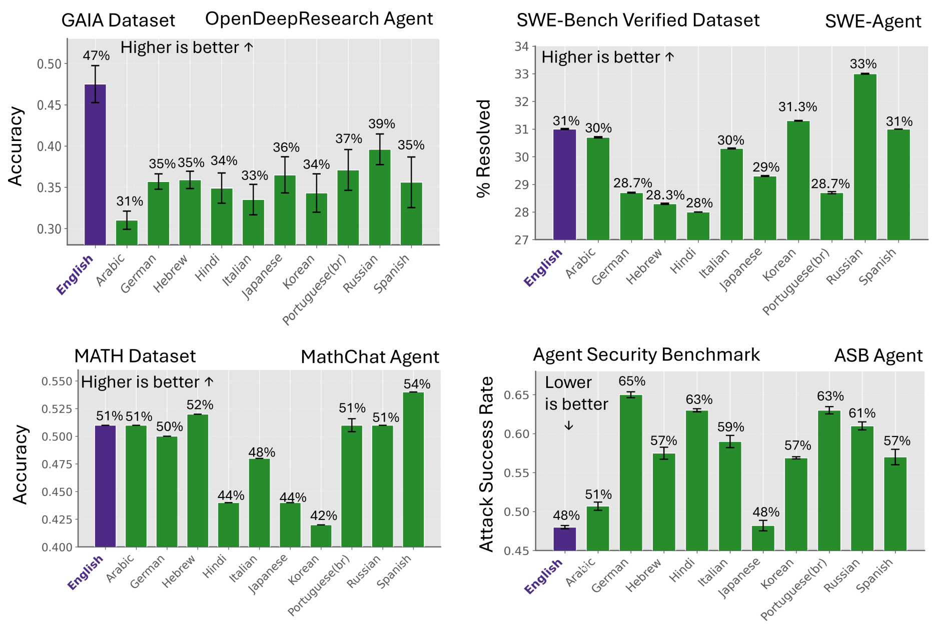

Figure 3: Performance of open-source agents across languages on four agentic benchmarks: GAIA, SWE-bench, MATH, and ASB. Each bar represents the agent's accuracy (or attack success rate in ASB) for a given language, with English shown as the baseline. Error bars indicate std across three runs. Performance differences reflect each agent's degradation or resilience in multilingual settings.

For GAIA, we used the OpenDeepResearch agent [von Platen et al., 2024], which integrates retrieval, web browsing, and tool use to support real-world reasoning. For MATH, we adopted MathChat [Wu et al., 2024], a zero-shot agent combining multi-turn reasoning with Python execution and the Wolfram Alpha tool. For SWE-bench, we applied SWE-agent [Yang et al., 2024b], which enables autonomous software reasoning through repository navigation, file editing, and test execution. For ASB, we built on the authors' existing infrastructure and evaluated the original ten-agent system against both direct and indirect prompt injection attacks across a variety of tasks and languages. Each agent was executed using the LLM backbone reported in its original implementation, all of which are considered multilingual models. Specifically, OpenDeepResearch used GPT o1, MATHChat used GPT-4, SWE-bench used GPT-4.1, and ASB used Qwen2. Full configuration details, including model versions and API providers, are provided in the Supplementary Materials.

Experiment Protocol. For each benchmark, the agent was evaluated three times in each of the 11 target languages, yielding a total of 33 runs per dataset. We report the mean and standard deviation over these runs. We used the original English task definitions and their translations, without modifying or translating internal agent logic and processing flows like system prompts or tools.

Metrics. We adopt the original evaluation metrics from each benchmark to ensure consistency with prior agent evaluations. For MathChat (Math), we report answer accuracy. For OpenDeepResearch (GAIA), we report the percentage of answers matching either the English or translated reference. For SWE-agent (SWE-bench), we report the percentage of resolved instances, defined as the percentage of submitted patches that successfully resolve the coding issue. For the ASB agent, we report the attack success rate (ASR), a standard metric in the security domain that represents the percentage of attacks that bypass the safety mechanisms. Additionally, we introduce a new metric: *Multilingual Effect*, which quantifies the performance or security gap between English and the average of all other languages. Given an evaluation metric M, the *Multilingual effect* is defined as follows:

$$\text{Multilingual Effect } = \frac{1}{n}\sum_{i=1}^{n} M_{\text{lang}_i} - M_{\text{en}} \tag{2}$$

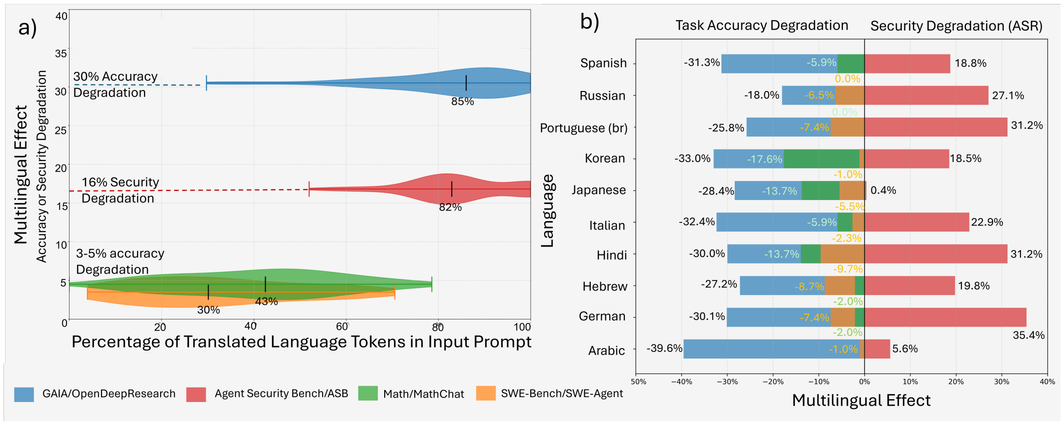

Figure 4: a) Multilingual effect as a function of the proportion of translated language tokens in input prompts. Each point represents a benchmark-agent pair, with the multilingual effect computed as the average relative degradation in performance or security across non-English languages. The trend suggests a correlation between input translation extent and multilingual vulnerability. b) Relative performance differences from English for each language, broken down by dataset. Negative values indicate a drop in performance compared to English, while positive values (notably in ASB) represent increased vulnerability. The trend highlights how multilingual effects vary by language and task type.

Where M_{en} denotes the performance in English, n is the number of non-English languages in the dataset (in our case $n = 10$), and M_{lang_i} represents the performance in the i-th non-English language.

4.2 Results

Figure 3 presents the performance of open-source agents across all four benchmarks in English and the ten target languages. In GAIA and ASB, we observe clear performance and security drops: non-English languages consistently underperformed compared to English, with reductions of up to 16% (GAIA) and a rise in vulnerability by up to 17% in ASB. Notably, SWE-bench and MATH exhibit only minor variation across languages, with most scores clustering around the English baseline.

These results reveal important differences in how multilingual degradation manifests across task types. Although all tasks require complex reasoning, some are more constrained than others. For instance, SWE-bench is limited to well-structured Python patches designed to fix specific test cases. As a result, the reliance on natural language explanations is reduced, with greater emphasis placed on strict Pythonic syntax and code correctness. On the other hand, GAIA focuses on solving real-world tasks with much more flexibility. Thus, the importance of the natural language problem statement is significantly higher. Additionally, in benchmarks like MATH and SWE-bench, the opportunity for translation is inherently limited, as a large portion of the input consists of mathematical notation or source code, thus naturally reducing the multilinguality effect. To understand this variation, we examine a potential driver: the proportion of localized, target-language-oriented input in each benchmark in each benchmark's input.

Interestingly, we observe that Japanese yields the lowest ASR (attack success rate) in the ASB benchmark, indicating the highest robustness to adversarial inputs. This result can be partially attributed to the fact that the ASB agent was implemented using the Qwen2 model [Yang et al., 2024a], which is known for its strong alignment for Japanese language tasks. Qwen2 has consistently demonstrated strong performance in Japanese-specific LLM benchmarks and leaderboards[7], suggesting that alignment and fine-tuning in a particular language can significantly enhance resilience against multilingual adversarial prompts. This reinforces the importance of language-specific alignment training in the development of robust and secure agentic systems.

Figure 4 examines the relationship between prompt composition and multilingual performance. Part (a) shows a correlation between the percentage of non-English tokens in the input and the average performance gap (relative to English) across all four datasets. Benchmarks with higher proportions

[7] https://rinnakk.github.io/research/benchmarks/lm/index.html

of localized, target-language-oriented input, such as GAIA and ASB, exhibit greater degradation, whereas SWE-bench, with predominantly English input (e.g., code), shows higher preservation.

From part (b), we can see that there is no clear correlation between multilingual security robustness (ASB) and multilingual performance degradation. This disconnect is especially clear in real-world, language-heavy tasks like GAIA, where performance drops sharply, while structured tasks like SWE-bench and MATH remain largely unaffected. This highlights that multilingual security alignment does not directly track with multilingual task accuracy, notably in language-rich agentic tasks.

5 Discussion

This section presents practical recommendations for multilingual agent deployment and directions for advancing the benchmark in future work.

5.1 Guidelines for Multilingual Evaluation and Risk Assessment

Language-Aware Deployment Guidelines. Before deploying an AI agent in a multilingual setting, analyze the linguistic composition of its expected input, particularly the balance between structured elements (e.g., code, formal queries) and localized natural language. Inputs with a high proportion of non-English content, especially those involving less formalized or more natural language, tend to increase the risk of performance and safety degradation. We therefore recommend that for any such case, developers should conduct a Multilingual Benchmark Assessment using a diverse, language-sensitive evaluation suite, such as ours, for AI agents operating across languages. This helps reveal hidden vulnerabilities and promotes reliable real-world behavior in multilingual conditions.

Prioritize Multilingual Adaptation by Task Type. Our findings suggest that the need for multilingual adaptation in agentic systems should be guided by task type. For structured tasks with minimal linguistic variability, such as coding, cross-lingual transfer can often be achieved with minimal adjustment. However, for complex, real-world tasks or safety-critical decisions (e.g., GAIA, ASB), multilingual robustness remains limited, and thus, dedicated multilingual alignment and adaptation are essential. MAPS offers a practical framework to identify where multilingual adaptation is needed, helping prioritize resource allocation for post-training based on task-specific language sensitivity.

Multilingual Inputs Amplify Agentic Security Vulnerabilities. Our evaluation on ASB revealed that multilingual adversarial inputs can bypass agent safety mechanisms with minimal sophistication. Direct translations of English jailbreak prompts—without any adaptation or obfuscation—were sufficient to induce policy-violating behavior in multiple languages. This highlights a critical risk: even simple adversarial prompts become significantly more effective when the input is localized, and are often sufficient to exploit security vulnerabilities in AI agents. Developers of safety-critical agentic systems should treat multilingual robustness as a core security concern and include translated prompts in safety evaluations using benchmarks like ours.

5.2 Benchmark Limitation

While MAPS represents the first multilingual suite for evaluating agentic AI systems, there are natural opportunities for future expansion. The current release includes four datasets, one agent per dataset, and ten target languages, offering a strong foundation for assessing multilingual robustness. Extending coverage to additional domains such as healthcare or legal reasoning, as well as incorporating multiple agents and extremely low-resource languages (e.g., Amharic or Uyghur), would further enhance the benchmark's scope and relevance. Nonetheless, the current suite already surfaces clear trends in performance and security degradation across languages, offering valuable insights for guiding multilingual deployment. We view this work as a meaningful starting point and invite the community to build on our open-source release to advance more inclusive and resilient agentic AI systems.

6 Conclusions

We introduce the first multilingual benchmark suite for evaluating agentic AI systems, addressing a critical gap in assessing language-specific performance and safety limitations. By adapting and localizing four widely used agentic benchmarks—GAIA, SWE-bench, MATH, and ASB—into ten diverse languages, our suite enables the analysis of agent behavior under multilingual conditions.

Constructed through a hybrid translation pipeline and human verification, the benchmark ensures high linguistic fidelity and structural consistency. Experimental results reveal consistent degradation in both performance and robustness when agents operate in non-English settings, particularly in tasks involving natural language reasoning and safety-critical behavior.

These findings underscore the importance of language-aware evaluation and targeted multilingual adaptation, especially for real-world agentic deployments. We view this benchmark as a practical and extensible foundation for building more inclusive, resilient, and globally reliable agentic AI systems, and we invite the community to build upon it.

References

A. A. Aakanksha, B. Ermis, S. Goldfarb-Tarrant, J. Kreutzer, M. Fadaee, S. Hooker, et al. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, 2024.

D. B. Acharya, K. Kuppan, and B. Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 2025.

D. I. Adelani, H. Liu, X. Shen, N. Vassilyev, J. O. Alabi, Y. Mao, H. Gao, and A. E.-S. Lee. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*, 2023.

R. Aharoni, M. Johnson, and O. Firat. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*, 2019.

M. Andriushchenko, A. Souly, M. Dziemian, D. Duenas, M. Lin, J. Wang, D. Hendrycks, A. Zou, Z. Kolter, M. Fredrikson, et al. Agentharm: A benchmark for measuring harmfulness of llm agents, 2024. *URL https://arxiv. org/abs/2410.09024*, 2024.

M. Chang, J. Zhang, Z. Zhu, C. Yang, Y. Yang, Y. Jin, Z. Lan, L. Kong, and J. He. Agentboard: An analytical evaluation board of multi-turn llm agents. *Advances in Neural Information Processing Systems*, 37:74325–74362, 2024.

L. Chua, B. Ghazi, Y. Huang, P. Kamath, R. Kumar, P. Manurangsi, A. Sinha, C. Xie, and C. Zhang. Crosslingual capabilities and knowledge barriers in multilingual large language models. *arXiv preprint arXiv:2406.16135*, 2024.

J. Dang, S. Singh, D. D'souza, A. Ahmadian, A. Salamanca, M. Smith, A. Peppin, S. Hong, M. Govindassamy, T. Zhao, S. Kublik, M. Amer, V. Aryabumi, J. A. Campos, Y.-C. Tan, T. Kocmi, F. Strub, N. Grinsztajn, Y. Flet-Berliac, A. Locatelli, H. Lin, D. Talupuru, B. Venkitesh, D. Cairuz, B. Yang, T. Chung, W.-Y. Ko, S. S. Shi, A. Shukayev, S. Bae, A. Piktus, R. Castagné, F. Cruz-Salinas, E. Kim, L. Crawhall-Stein, A. Morisot, S. Roy, P. Blunsom, I. Zhang, A. Gomez, N. Frosst, M. Fadaee, B. Ermis, A. Üstün, and S. Hooker. Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL `https://arxiv.org/abs/2412.04261`.

Y. Deng, W. Zhang, S. J. Pan, and L. Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

L. Gioacchini, G. Siracusano, D. Sanvito, K. Gashteovski, D. Friede, R. Bifulco, and C. Lawrence. Agentquest: A modular benchmark framework to measure progress and improve llm agents. *arXiv preprint arXiv:2404.06411*, 2024.

N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, and A. Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint 2407.21783*, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023.

J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International conference on machine learning*, pages 4411–4421. PMLR, 2020.

B. Jha. The babel effect: Analyzing multilingual performane discrepancies in large language models. *Eng Appl Sci J*, 1(1):1–4, 2024.

C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

D. Ki and M. Carpuat. Guiding large language models to post-edit machine translation with error annotations. *arXiv preprint arXiv:2404.07851*, 2024.

P. Koehn and R. Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, 2017.

Y. Liang, N. Duan, Y. Gong, N. Wu, F. Guo, W. Qi, M. Gong, L. Shou, D. Jiang, G. Cao, et al. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*, 2020.

G. Mialon, C. Fourrier, T. Wolf, Y. LeCun, and T. Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

T. A. Mohammed. From google translate to chatgpt: The use of large language models in translating, editing, and revising. In *Role of AI in Translation and Interpretation*, pages 1–32. IGI Global Scientific Publishing, 2025.

F. Shi, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, and J. Wei. Language models are multilingual chain-of-thought reasoners, 2022.

H. Trivedi, T. Khot, M. Hartmann, R. Manku, V. Dong, E. Li, S. Gupta, A. Sabharwal, and N. Balasubramanian. Appworld: A controllable world of apps and people for benchmarking interactive coding agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16022–16076, 2024.

P. von Platen, Y. Wang, T. Le Scao, and H. F. team. Open deep research: A blueprint for transparent foundation model science. `https://huggingface.co/blog/open-deep-research`, 2024. URL `https://huggingface.co/blog/open-deep-research`. Hugging Face Blog.

W. Wang, Z. Tu, C. Chen, Y. Yuan, J.-t. Huang, W. Jiao, and M. R. Lyu. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*, 2023.

Y. Wu, F. Jia, S. Zhang, H. Li, E. Zhu, Y. Wang, Y. T. Lee, R. Peng, Q. Wu, and C. Wang. Mathchat: Converse to tackle challenging math problems with llm agents. *arXiv preprint arXiv:2306.01337*, 2023.

Y. Wu, F. Jia, S. Zhang, H. Li, E. Zhu, Y. Wang, Y. T. Lee, R. Peng, Q. Wu, and C. Wang. Mathchat: Converse to tackle challenging math problems with llm agents. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

J. Xie, K. Zhang, J. Chen, T. Zhu, R. Lou, Y. Tian, Y. Xiao, and Y. Su. Travelplanner: A benchmark for real-world planning with language agents. In *International Conference on Machine Learning*, pages 54590–54613. PMLR, 2024.

F. F. Xu, Y. Song, B. Li, Y. Tang, K. Jain, M. Bao, Z. Z. Wang, X. Zhou, Z. Guo, M. Cao, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

B. Xue, H. Wang, R. Wang, S. Wang, Z. Wang, Y. Du, B. Liang, and K.-F. Wong. Mlingconf: A comprehensive study of multilingual confidence estimation on large language models. *arXiv preprint arXiv:2410.12478*, 2024.

J. Yan, P. Yan, Y. Chen, J. Li, X. Zhu, and Y. Zhang. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*, 2024.

A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, and Z. Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

J. Yang, C. Jimenez, A. Wettig, K. Lieret, S. Yao, K. Narasimhan, and O. Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37: 50528–50652, 2024b.

S. Yao, N. Shinn, P. Razavi, and K. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains, 2024. *URL https://arxiv. org/abs/2406.12045*, 2024.

A. Yehudai, L. Eden, A. Li, G. Uziel, Y. Zhao, R. Bar-Haim, A. Cohan, and M. Shmueli-Scheuer. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*, 2025.

O. Yoran, S. J. Amouyal, C. Malaviya, B. Bogin, O. Press, and J. Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? *arXiv preprint arXiv:2407.15711*, 2024.

H. Zhang, J. Huang, K. Mei, Y. Yao, Z. Wang, C. Zhan, H. Wang, and Y. Zhang. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. *arXiv preprint arXiv:2410.02644*, 2024.