Dataset details

Multilingual Benchmark for Agent Performance and Security (MAPS)

This is the first Multilingual Agentic AI Benchmark for evaluating agentic AI systems across different languages and diverse tasks. Benchmark enables systematic analysis of how agents perform under multilingual conditions. To balance performance and safety 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 safety-relevant prompts from ASB. Each task was translated into 11 target languages resulting in a total of 9.6K multilingual tasks.

Dataset Description

This benchmark is designed to evaluate agentic AI systems for both performance and safety across a wide range of tasks in a multilingual setting. It enables testing how well agents perform when operating in different languages, covering realistic tasks from multiple domains:

GAIA: Web search and tool-use tasks that test an agent's ability to interact with external tools and follow multi-step reasoning.

MATH: Complex mathematical problem-solving tasks from seven topics, requiring structured reasoning and accurate computation.

SWE-bench: Software engineering tasks involving real-world GitHub issues, focusing on code understanding, bug fixing, and technical reasoning.

ASB (Agent Security Benchmark): Safety-focused tasks designed to probe agent behavior under adversarial or sensitive scenarios, ensuring safe and aligned outputs across languages.

languages

Each task in the benchmark is translated into the following 11 languages to enable comprehensive multilingual evaluation: Spanish (es), German (de), Arabic (ar), Russian (ru), Japanese (ja), Portuguese (pt), Hindi (hi), Hebrew (he), Korean (Ko), Italian (it), Chinese (zh)

Dataset Size

Each dataset in the benchmark includes a fixed number of instances per language, all translated into 11 languages. Below is the breakdown (including english):

- GAIA: 165 tasks per language × 12 languages = 1,980 tasks total
- MATH: 140 tasks per language × 12 languages = 1,680 tasks total
- SWE-bench: 100 tasks per language × 12 languages = 1,200 tasks total
- ASB: 400 attack per language × 12 languages = 4,800 attacks total

Direct Use

- **Compare multilingual robustness across agent designs or toolchains:** Evaluate how different agent architectures, prompting strategies, or tool-use capabilities perform across languages. This helps identify which designs are more robust to linguistic variation in task execution.
- **Stress test agents for safe behavior in non-English inputs:** Use the Agent Security Benchmark (ASB) subset to probe safety risks in multiple languages. This scenario reveals whether agents behave safely and consistently when faced with adversarial or sensitive prompts beyond English.
- **Benchmark cross-lingual generalization in reasoning, code, and safety tasks:** Assess agents on their ability to generalize core reasoning, coding, and safety principles across languages using datasets like GAIA, MATH, SWE-bench, and ASB.

- **Analyze performance drop-offs or safety regressions across languages:** Track how performance or safety behavior degrades in certain languages compared to English. This helps uncover biases, translation artifacts, or limitations in the agent's multilingual handling.

Data Splits

Users can filter the benchmark tasks using two main criteria: by dataset (e.g., GAIA, MATH, SWE-bench, ASB) and by language (from the 11 supported languages). This flexible filtering enables targeted evaluation of agent performance and safety across specific domains and languages.

Data format

All datasets are available in json format.

Curation Rationale

To build our multilingual benchmark, we use a hybrid machine–generation and human–verification pipeline. AI-based processing produces language variants at scale, while native speakers verify meaning and nuance. Each task is represented consistently across the eleven diverse languages, ensuring faithful intent preservation and enabling reliable cross-language evaluation.

Data Collection and Processing

Original English data is collected from below mentioned citations [1,2,3,4].

We adopt a hybrid multi-stage translation pipeline that systematically combines the format-preserving strengths of Machine translation with the contextual refinement capabilities of LLMs, followed by manual verification for quality assurance. More details about the hybrid translation pipeline are available in our Research Paper.

Annotations

Each item was independently rated by a bilingual annotator fluent in English and the target language Annotators evaluated three criteria on a 1~5 Likert scale: adequacy (semantic fidelity), fluency (grammatical and stylistic naturalness), and formatting accuracy (preservation of special elements such as LaTeX, variable names, and code). A final metric, answerability, captured whether the translation preserved the original intent well enough for the annotator to confidently answer the question as if it were posed in English. More details about the Annotations are available in our Research Paper.

Users should be made aware of the risks, biases and limitations of the dataset.

Citation

1. Mialon, G., Fourrier, C., Wolf, T., LeCun, Y., & Scialom, T. (2023). GAIA: A Benchmark for General-AI Assistants. *ICLR 2023*. https://openreview.net/forum?id=GAIA2023
2. Zhang, H., Huang, J., Mei, K., Yao, Y., Wang, Z., Zhan, C., Wang, H., & Zhang, Y. (2024). Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents. *arXiv 2410.02644*. https://arxiv.org/abs/2410.02644
3. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., & Steinhardt, J. (2021). Measuring Mathematical Problem Solving with the MATH Dataset. *arXiv 2103.03874*. https://arxiv.org/abs/2103.03874
4. Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., & Narasimhan, K. (2023). SWE-Bench: Can Language Models Resolve Real-World GitHub Issues? *arXiv 2310.06770*. https://arxiv.org/abs/2310.06770