GAIA-v2-LILT: Multilingual Adaptation of Agent Benchmark beyond Translation

Yunsu Kim* **Kaden Uhlig*** **Joern Wuebker**
LILT
{yunsu.kim,kaden.uhlig,joern}@lilt.com

Abstract

Agent benchmarks remain largely English-centric, while their multilingual versions are often built with machine translation (MT) and limited post-editing. We argue that, for agentic tasks, this minimal workflow can easily break benchmark validity through query-answer misalignment or culturally off-target context. We propose a refined workflow for adapting English benchmarks into multiple languages with explicit functional alignment, cultural alignment, and difficulty calibration using both automated checks and human review. Using this workflow, we introduce *GAIA-v2-LILT*, a re-audited multilingual extension of GAIA covering five non-English languages. In experiments, our workflow improves agent success rates by up to 32.7% over minimally translated versions, bringing the closest audited setting to within 3.1% of English performance while large gaps remain in many other cases. This indicates that a substantial share of the multilingual performance gap is benchmark-induced measurement error, motivating task-level alignment when adapting English benchmarks across languages. The data is available as part of the MAPS package.[1] We also release the code used in our experiments.[2]

1 Introduction

Modern AI systems are moving from single-turn assistants to autonomous agents that perform multi-step reasoning with external tools (Mialon et al., 2023a; Wang et al., 2024; Xi et al., 2025). However, agent benchmarks remain largely English-centric (Yao et al., 2022; Deng et al., 2023; Liu et al., 2024; Zhou et al., 2024; Jimenez et al., 2024; Mialon et al., 2023b; Barres et al., 2025; Wei et al., 2025; Xie et al., 2024; Patwardhan et al., 2025). This limits its reliable measurement for non-English users and reinforces an English bias in training and optimization of a large language model (LLM) (Ahuja et al., 2023; Zhang et al., 2023; Nguyen et al., 2024).

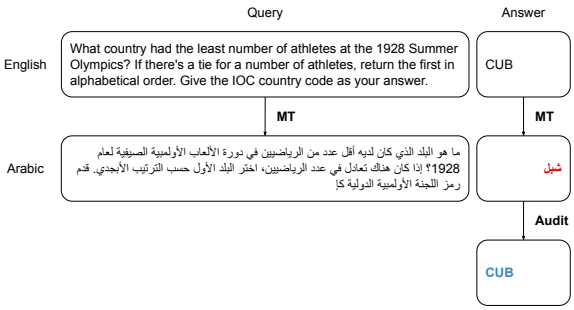

Figure 1: Example of query-answer misalignment by over-translation. MT violates the required answer format, which is subsequently corrected during the human audit.

A common workaround is to machine-translate an English dataset and apply basic post-editing to a limited subset (Hofman et al., 2025; Wang et al., 2025b; Issaka et al., 2026). This approach is standard for conventional document translation, but in agent benchmarks it can fail to preserve task mechanics. In Figure 1, the answer was incorrectly translated into the Arabic word for "lion cub", interpreting the IOC country code "CUB" (for Cuba) as a common noun. Not only is the translation semantically irrelevant, but it also violates the query's constraint to provide a three-letter country code.

Including such functional misalignment, this paper analyzes common pitfalls in translation-based workflows for adapting English agent benchmarks into other languages and proposes practical countermeasures. Our contributions are as follows:

- We show how conventional translation quality issues can compromise the integrity of agent tasks.
- We identify specific issues that affects the validity and appropriateness of multilingual agent benchmarking: functional alignment,

*Equal contribution.
[1] https://huggingface.co/datasets/Fujitsu-FRE/MAPS/viewer/GAIA-v2-LILT
[2] https://github.com/lilt/gaia-v2-lilt

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 140–148

cultural alignment, and difficulty calibration.

- We propose a refined workflow for addressing these issues, combining automatic checks and targeted human review.
- We apply this workflow to build the GAIA-v2-LILT dataset and show that correcting these issues materially changes measured agent performance.

2 Related Work

Multilingual Agent Benchmarks Existing multilingual agent benchmarks span general reasoning (Hofman et al., 2025), function calling (Kulkarni et al., 2025; Luo et al., 2026), coding (Raihan et al., 2025), web navigation (Wang et al., 2025b), and computer use (Yang et al., 2025). They rely on MT from English with minimal or no human edits, whereas our work introduces dedicated alignment procedures tailored for agentic tasks.

Some recent works build multilingual agent tasks from scratch using native components (Almeida et al., 2025; Kautsar et al., 2025; Zhou et al., 2025). While ideal for capturing localized nuances, native creation is costly and labor-intensive to set up for each language. This work focuses on adapting well-established English benchmarks for standardized evaluation and scalability.

MT Post-Editing Conventional MT post-editing (Balling et al., 2014; Koponen, 2016; Vieira, 2019) primarily prioritizes fluency and adequacy of the translated text, usually based on established error types (Lommel et al., 2013). Recent studies show that reviewers need additional context around the translation for fixing document-level errors, such as inconsistent terms or unclear pronouns (Agrawal et al., 2024; Koneru et al., 2024; Castilho and Knowles, 2025).

While these methods focus on linguistic correctness and meaning preservation, our approach goes beyond surface-level language fixes to ensure functional alignment, cultural relevance, and consistent task difficulty so the benchmarks remain executable and solvable.

3 Task Definition

We explain the basics of agentic problem solving to clarify our data curation and benchmarking.

Task In this paper, an agentic task is a user query that requires at least one external *action* beyond internal language modeling. Typical actions include web search, page parsing, file inspection, numerical computation, and scripting. A successful agent coordinates these steps to return an answer matching the exact output specification. This differs from standard question answering because the model must plan and execute a procedure rather than just write a fluent response.

Properties Following the GAIA dataset (Mialon et al., 2023b), our tasks feature:

- Single-turn interactions.
- Queries often have strict formatting constraints for the output, such as digit precision or list order.
- Text-based expected answers to enable deterministic grading.

Solving Process Agent's execution involves four stages: query comprehension, plan construction, tool execution, and response synthesis. Agents allocate a finite budget across these steps; If a query is malformed or culturally misaligned, the agent might waste this budget fixing instructions instead of solving the task.

Evaluation Most setups measure accuracy on the final answer using an exact match. To focus on reasoning, we allow minor variations by removing spaces, lowercasing, and stripping punctuation (Mialon et al., 2023b). While practical, this binary metric ignores the intermediate steps; an agent might reason correctly but fail due to a flawed answer key or an ill-formed query.

4 Multilingual Adaptation

We explain how to adapt an English task (from Section 3) into another language. For each step, we detail the motivation and provide examples of common issues.

4.1 Translation

The first step translates the English data into the target language. This aims to maintain the dataset size and topic coverage while ensuring the invariance of task function across languages. Typically, MT performs the first pass, but it introduces two main classes of problems.

Translationese Linguistic artifacts arise from literal translations by MT that create *translationese*, which deviates from natural patterns of the target language. It typically manifests through retaining

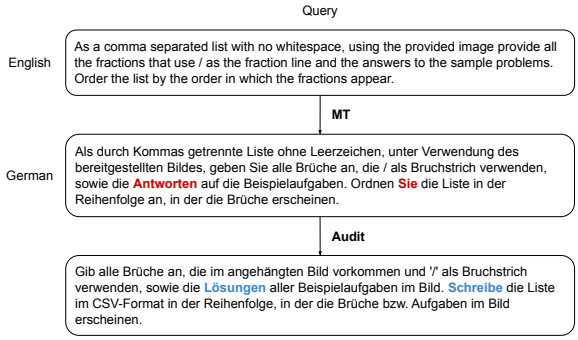

Figure 2: Example of correcting translationese. The audit refines the literal translation to ensure natural phrasing, correct terminology, and appropriate formality.

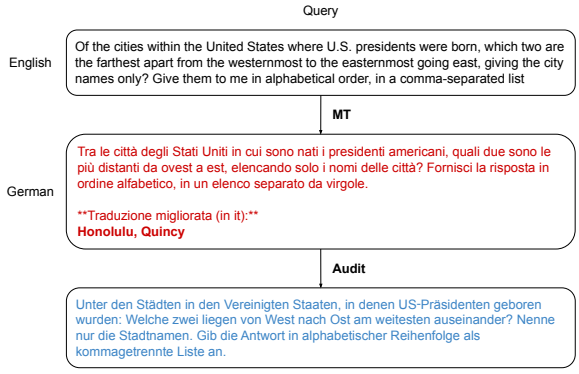

Figure 3: Example of correcting hallucination. Instead of the target German, the MT generates Italian and hallucinates the actual answer within the text. The audit restores the correct language and removes the answer.

word order of the source language, word-for-word conversion of idioms, incorrect formality or honorifics (Koppel and Ordan, 2011; Li et al., 2025a).

The resulting data carries an inherently artificial quality, as shown in the example in Figure 2. Here, *Antworten* is used for mathematical problems where *Lösungen* is the standard term. Also, it uses the formal pronoun *Sie*, which is inappropriate for usual human-AI interaction. The corrected text ensures task realism through native syntax, proper formality, and idiomatic phrasing.

These patterns are especially prominent in modern models trained on vast amounts of synthetic data. In an agentic context, this forces the model to decode unnatural phrasing rather than focus on high-level reasoning, compromising its planning and execution logic (Hofman et al., 2025; Wang et al., 2025b).

Hallucination The MT field is transitioning from neural machine translation (NMT) to LLMs for improved contextual fluency. While LLMs produce more natural prose, they may introduce generative artifacts beyond linguistic errors such as outputting a completely wrong language, including a hint or the gold answer in the prompt (Huang et al., 2025).

Such artifacts cause execution failures or shortcuts, shifting the benchmark from a test of intelligence to a test of robustness against corrupted logic. In Figure 3, the model generated text in Italian rather than the requested German. Furthermore, it answered the prompt during translation ("*Honolulu, Quincy*").

4.2 Functional Alignment

While standard post-editing can fix the translation issues above, it does not guarantee task solvability. A critical failure mode is query-answer mismatch, where the translated answer no longer aligns with the translated prompt. We divide this into two cases:

Under-Translation This occurs when a reference answer remains in English even though the context requires a localized term, causing false negatives. If an agent outputs the correct local name but the evaluator expects the English original, valid reasoning is unfairly penalized. For example, a German chess task will incorrectly reject the valid local notation *Td5* ("Turm") if the answer key is left as the English *Rd5* ("Rook").

Over-Translation Conversely, some answers must remain in English to maintain task integrity. If a task requires extracting a specific string from an English document, translating that string prevents exact matching (Figure 1). Similarly, transliterating obscure entities that lack established local conventions often introduces extra ambiguity.

Distinguishing between these cases is a delicate nuance problem. It requires reviewers to balance deep cultural conventions with a technical understanding of the question type to determine whether localization aids or obstructs the evaluation logic.

4.3 Cultural Alignment

Even when tasks remain solvable, retaining assumptions in the source region can make them unnatural for local users. These include geography, policy logic, measurement units, platform conventions, and social norms.

In Figure 4, a query about returning bottles in a road trip was translated into Korean but remained anchored to U.S. highways and imperial units. The

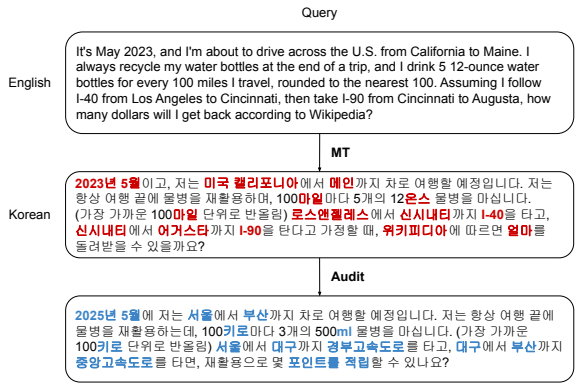

Figure 4: Example of cultural alignment. The MT retains U.S. geography and imperial units. The audit localizes the query with Korean routes, measurements, and a reward system. The date is also updated to align with the actual availability of recycling rewards in South Korea.

corrected version replaced route references, converted units, updated the date context, and switched to a Korean recycling incentive system (where no local equivalent exists, the literal U.S. context is retained).

While literal translations represent realistic scenarios for expats, full localization better reflects the authentic usage of the primary target population. Crucially, leaving queries as literal translations often prompts agents to simply translate the text back to English and solve the original U.S.-centric task, bypassing true multilingual reasoning (Li et al., 2025b; Wang et al., 2025a).

4.4 Difficulty Calibration

Beyond linguistic, functional, and cultural alignment, converting a task can still alter its inherent difficulty. For instance, a query becomes significantly harder if relevant information is scarce on the web in the target language, or if localized currencies complicate simple math. Therefore, reviewers must manually verify the solving process to ensure the difficulty aligns with the English original, e.g., by conducting local web searches themselves.

Figure 5 demonstrates difficulty calibration using a word riddle. While the draft is fluent and preserves the main task function, it lowers the reading difficulty by writing the sentence normally around the English clue "*tfel*". The corrected version restores the original challenge by fully reversing the Korean sentence and localizing the clue word.

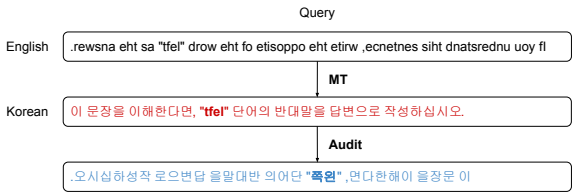

Figure 5: Example of difficulty calibration.

5 GAIA-v2-LILT

Considering all dimensions of Section 4, we constructed a multilingual agent benchmark *GAIA-v2-LILT*, based on the machine-translated version of GAIA (Mialon et al., 2023b) by MAPS (Hofman et al., 2025). GAIA-v2-LILT contains 165 query-answer pairs (validation set) for each of the five non-English languages: Arabic, German, Hindi, Korean, and Portuguese (Brazil).

We built a hybrid review workflow combining automatic checks with specialized human auditing to effectively correct the aforementioned issues (Figure 6). It is specifically designed to counter two major evaluation pitfalls:

- LLM's *self-preference*: models favoring text similar to their own outputs (Zheng et al., 2023)
- Human's *fluency bias*: reviewers mistakenly assuming that well-written text is technically correct (Gudibande et al., 2024)

Deterministic Filtering Before applying model or human judgment, we used fast, rule-based scripts to catch high-impact, objective defects. Specifically, these scripts check:

- whether the translation is in the target language (language identification)
- whether the expected answer is leaked in the query (string matching)
- whether all fixed term categories are preserved, e.g., numbers, URLs, etc. (placeholder recall)

Relying on exact matching rather than generative evaluation, this step does not have the self-preference problem of LLMs.

LLM Judges Next, we used LLMs to pre-evaluate the deeper semantic and functional aspects. Rather than asking for a single holistic score, we decomposed the review into narrow, single-axis

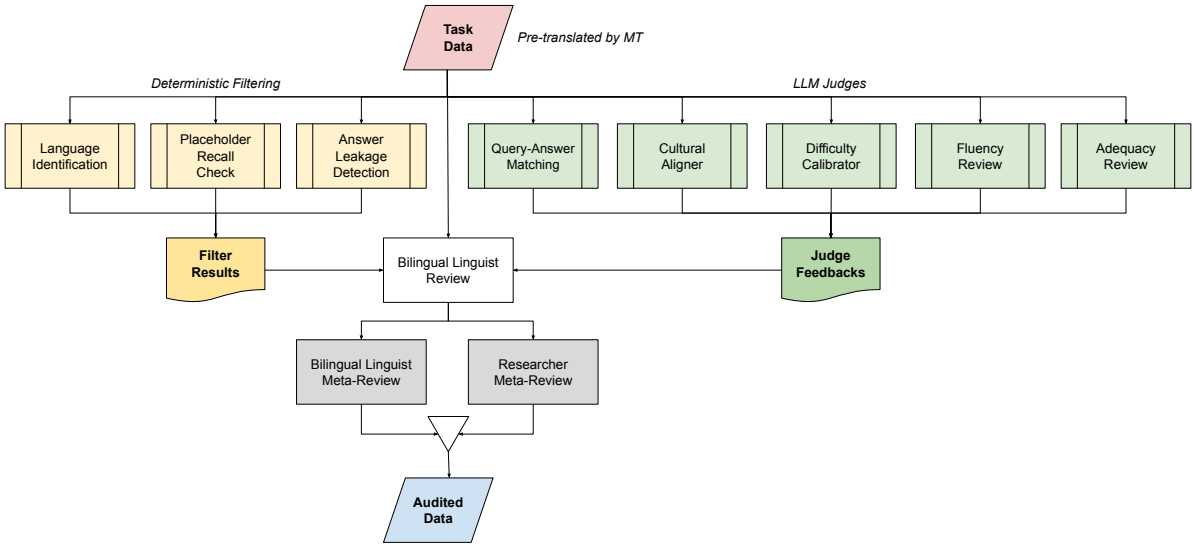

Figure 6: Review workflow for constructing GAIA-v2-LILT.

judgments. This design helps limit self-preference bias and can better track human judgments, because each decision is made over a short context and a concrete question instead of requiring a deep global analysis of the text (Saha et al., 2024; Feng et al., 2025; Lee et al., 2025). Accordingly, separate judges independently evaluated fluency, adequacy, query-answer compatibility, and cultural appropriateness using binary labels.

Human Audit Finally, we employed bilingual human linguists to review the pre-translated tasks. We conducted 1-on-1 training calls with each reviewer to detail the alignment issues from Section 4, placing strict emphasis on validating functional integrity over mere fluency.

During the audit, results from the deterministic filters and LLM judges were displayed alongside the task text. This highlights critical issues immediately, helping reviewers prioritize their time budget toward the most difficult cases. Reviewers also had to mark explicit checkboxes for whether each issue category applied to the task. These structured steps help reduce human fluency bias.

Each task was reviewed by one linguist and meta-reviewed by another linguist and a machine learning researcher. As shown in Table 1, the final edit rates against MAPS are substantial, underscoring the necessity of this extensive human audit.

Language	Task	Word	Char
Arabic	84.8	25.4	19.4
German	81.2	30.0	22.2
Hindi	100.0	55.4	46.3
Korean	92.1	36.9	27.9
Portuguese	87.9	25.0	19.5

Table 1: Edit rates [%] of GAIA-v2-LILT against the original MAPS version of GAIA.

6 Experiments

With the benchmark tasks rigorously verified, we evaluated frontier LLMs on the resulting GAIA-v2-LILT dataset to measure the true impact of our review workflow.

Setup We evaluated three leading models (GPT-5.4, Gemini 3.1 Pro, Claude Opus 4.6) using the Open Deep Research agent (Roucher et al., 2025). The agents were equipped with web search (Exa[3]), speech recognition, image captioning, and file read tools. For each task, we limited the manager agent to 12 steps and the search subagent to 20 steps.

Impact of Audits Table 2 compares model performance on the raw MT drafts versus the human-corrected final data. Before correction, multilingual scores lag substantially behind the English baselines. However, after human audit, performance across all five target languages improves dramatically, yielding absolute gains ranging from +10.9% to +32.7%. This recovery reduces the performance gap with English to a minimum of 3.1%, though Arabic still shows a larger gap of up to 30.3%.

These results demonstrate that MT may mask an agent's true capabilities significantly. Our au-

[3] https://exa.ai

	Arabic		German		Hindi		Korean		Portuguese		
Model	MT	Audit	MT	Audit	MT	Audit	MT	Audit	MT	Audit	English
GPT-5.4	32.1	47.3	47.3	63.6	34.6	60.0	33.3	62.4	47.3	58.2	66.7
Gemini 3.1 Pro	34.6	52.1	49.7	66.7	38.2	63.6	34.6	64.8	48.5	65.5	73.9
Claude Opus 4.6	32.1	49.1	49.7	66.7	29.7	62.4	33.3	58.8	49.1	63.0	79.4

Table 2: Agent performance before and after correction (pass@1 accuracy [%]).

dits reveal the actual performance trends across languages, proving that carefully aligned data is essential for accurate multilingual benchmarking of agents.

Analysis of Edits To understand how the quality issues discussed in Section 4 affect benchmark evaluation, we compare model outputs on the original and corrected versions of the same tasks. For each task, we check whether linguists flagged a specific issue category during the audit. We then measure whether correcting that issue caused the model's correctness to flip, changing the outcome from wrong to right or vice versa.

Figure 7 shows the analysis. Since multiple issues can co-occur on the same task, flip rates are correlational rather than causal. Nevertheless, functional alignment stands out with the highest flip rate (67.9%), consistent with the expectation that a mistranslated gold answer penalizes the model regardless of actual performance. Cultural alignment, adequacy, and difficulty calibration flip outcomes roughly half the time, indicating strong association with correctness judgments. Notably, despite being the most frequently flagged, fluency issues have the lowest flip rate (40.6%).

This suggests that stylistic perfection is weakly associated with evaluation outcome changes. Consequently, resource-constrained localization efforts should prioritize functional and cultural fidelity to maximize the integrity of the evaluation.

7 Conclusion

Building on MAPS (Hofman et al., 2025), we created the GAIA-v2-LILT dataset to directly address the task-logic failures caused by MT in agent benchmarks. We developed a rigorous review workflow combining deterministic filters, single-axis LLM judges, and structured human audits to mitigate both LLM self-preference and human fluency bias.

By correcting functional and cultural misalignments while preserving task difficulty, models improved accuracy by up to 32.7%. This highlights the critical, yet often overlooked, impact of non-linguistic alignment errors when translating agent benchmarks. We argue that future multilingual agent evaluations must adopt similar workflows to ensure accurate measurement.

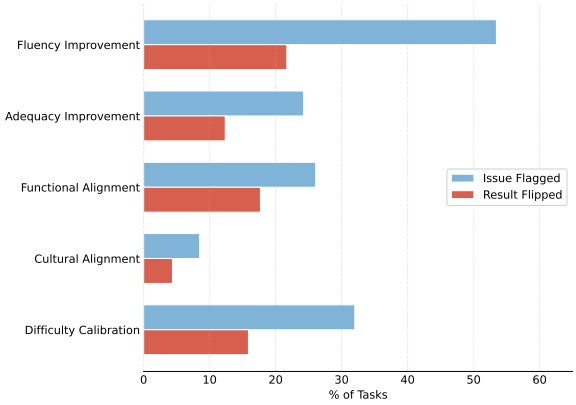

Figure 7: Flagging rates of issues during the audit and result flipping rates (Gemini 3.1 Pro) among the flagged tasks. Aggregated across all languages.

Limitations

Our review process omits baseline agent testing which can detect hidden technical flaws, e.g., search APIs failing to retrieve required information (due to poor retrieval or geo-blocking) or models bypassing intended tool usage by relying on memorized training data. Since configuring execution environments is impractical for non-technical annotators, we leave the integration of automated agent testing into the review loop via annotator-friendly interfaces to future work.

Also, file attachments (e.g., images, audio, and PDFs) were not localized, as building a consistent editing pipeline for diverse asset types is highly complex. While keeping these files in English does not break the core task logic, it prevents perfect localization. The development of streamlined multimodal localization workflows is left to future work.

Acknowledgments

We sincerely thank the human annotators for reviewing the GAIA-v2-LILT dataset, and the authors of MAPS (Hofman et al., 2025) for coordinating the data release.

References

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André FT Martins. 2024. Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Thales Sales Almeida, João Guilherme Alves Santos, Thiago Laitz, and Giovana Kerche Bonás. 2025. Ticket-bench: A kickoff for multilingual and regionalized agent evaluation. *arXiv preprint arXiv:2509.14477*.

Laura Winther Balling, Michael Carl, and Sharon O'Brian. 2014. *Post-editing of machine translation: Processes and applications*. Cambridge Scholars Publishing.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. 2025. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. *arXiv preprint arXiv:2506.07982*.

Sheila Castilho and Rebecca Knowles. 2025. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4):986–1016.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems*, volume 36, pages 28091–28114. Curran Associates, Inc. NeurIPS 2023 Datasets and Benchmarks Track.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-mad: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107.

Arnav Gudibande, Eric Wallace, Charlie Victor Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2024. The false promise of imitating proprietary language models. In *The Twelfth International Conference on Learning Representations*.

Omer Hofman, Jonathan Brokman, Oren Rachmil, Shamik Bose, Vikas Pahuja, Toshiya Shimizu, Trisha Starostina, Kelly Marchisio, Seraphina Goldfarb-Tarrant, and Roman Vainshtein. 2025. Maps: A multilingual benchmark for agent performance and security. *arXiv preprint arXiv:2505.15935*. Accepted to Findings of EACL 2026.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Sheriff Issaka, Erick Rosas Gonzalez, Lieqi Liu, Evans Kofi Agyei, Lucas Bandarkar, Nanyun Peng, David Ifeoluwa Adelani, Francisco Guzmán, and Saadia Gabriel. 2026. Translation as a scalable proxy for multilingual evaluation. *arXiv preprint arXiv:2601.11778*.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. SWE-bench: Can language models resolve real-world GitHub issues? In *The Twelfth International Conference on Learning Representations*. Oral.

Muhammad Dehan Al Kautsar, Aswin Candra, Muhammad Alif Al Hakim, Maxalmina Satria Kahfi, Fajri Koto, Alham Fikri Aji, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Genta Indra Winata. 2025. Seadialogues: A multilingual culturally grounded multi-turn dialogue dataset on southeast asian languages. *arXiv preprint arXiv:2508.07069*.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725.

Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, (25):131–148.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326.

Vaibhav Kulkarni, Tushar Mhetre, Amit Singh, Ripan Saha, Saurav Chatterjee, Sreevijay Raj, Sandesh Swamy, Shachi Dave, Raghavan Srinivasan, Diptanu Das, and Venu Govindaraju. 2025. MASSIVE-agents: A benchmark for multilingual function calling in 52 languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9309–9342, Suzhou, China. Association for Computational Linguistics.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025a. Lost in literalism: How supervised training shapes translationese in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025b. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*. Poster.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*.

Zheng Luo, T Pranav Kutralingam, Ogochukwu N Okoani, Wanpeng Xu, Hua Wei, and Xiyang Hu. 2026. Lost in execution: On the multilingual robustness of tool calling in large language models. *arXiv preprint arXiv:2601.05366*.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023a. Augmented language models: a survey. *Transactions on Machine Learning Research*.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023b. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. 2025. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *arXiv preprint arXiv:2510.04374*.

Md Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025. 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)*, pages 11432–11461.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. 'smolagents': a smol library to build great agentic systems. `https://github.com/huggingface/smolagents`.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370.

Lucas Nunes Vieira. 2019. Post-editing of machine translation. In *The Routledge handbook of translation and technology*, pages 319–336. Routledge.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18:186345.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025a. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5075–5094.

Peng Wang, Chunping Tao, Lu Niu, Hao Jiang, Shuai Qiao, Ji Xiang, Jiawei Han, Haoran Xu, Chao Xu,

Tairan Ge, Zhaopeng Tu, Wayne Xin Zhao, and Ji-Rong Wen. 2025b. X-webagentbench: Evaluating multilingual llm agents on multimodal interactive web navigation tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19097–19122, Vienna, Austria. Association for Computational Linguistics.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 9 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, pages 52040–52094. Curran Associates, Inc. NeurIPS 2024 Datasets and Benchmarks Track.

Pei Yang, Hai Ci, and Mike Zheng Shou. 2025. macOSWorld: A multilingual interactive benchmark for GUI agents. *arXiv preprint arXiv:2506.04135*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc. NeurIPS 2022 Main Conference Track.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in english: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, and 1 others. 2025. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*. Poster.