Maalim
e-ISSN: 2600-6944 | p-ISSN: 2170-0052
Volume 19 Issue 1/ Juin 2026 PP. 11-24

Translation in the Age of AI Agents

Dr. Fouad Bousetouane *
The University of Chicago, USA
f.bousetouane@gmail.com

DOI: 10.33705/1111-019-001-001

Received: 28/02/2026
Accepted: 18/05/2026
Published: 30/06/2026

***Corresponding Author**

Citation :
Bousetouane,F. (2026).
Translation in the Age of AI Agents
Maalim
I(1), 11-24

Abstract:

Translation is entering a new technological phase. For decades, machine translation systems converted sentences from one language to another with increasing fluency, yet they remained constrained by context loss, domain ambiguity, cultural nuance, and limited mechanisms for verification. The emergence of large language models (LLMs) and AI agents changes the translation function from a passive text-conversion task into an active, tool-using, context-aware workflow. An AI translation agent can retrieve domain knowledge, consult glossaries, check terminology, preserve institutional tone, compare alternative translations, evaluate uncertainty, and request human review when risk is high. This paper argues that agentic translation represents a paradigm shift in diplomatic, scientific, technical, legal, healthcare, and enterprise communication. The shift is especially important because the web is linguistically imbalanced: English dominates online content, while French and Arabic occupy much smaller shares, shaping the available knowledge base for AI systems. This imbalance creates both an opportunity and a risk. Agentic translation can bridge knowledge gaps by grounding translation in trusted corpora and domain-specific references, but it must be evaluated through human-in-the-loop processes and automated adversarial evaluation frameworks such as ProofAgent Harness framework. The paper proposes a reference architecture for agentic translation, identifies key domain use cases, and outlines evaluation dimensions for reliable deployment.

Keywords: AI agents; machine translation; large language models; agentic translation; domain grounding; human-in-the-loop; ProofAgent Harness.

الترجمة في عصر الذكاء الاصطناعي

د. فؤاد بوسطوان

جامعة شيكاكو، الولايات المتحدة الأمريكية

الملخص:

تشهد وظيفة الترجمة تحولا جوهريا مع ظهور النماذج اللغوية الكبيرة ووكلاء الذكاء الاصطناعي .فبعد أن كانت الترجمة الآلية تركز أساسا على تحويل الجمل من لغة إلى أخرى، أصبحت اليوم قادرة على أن تتحول إلى سير عمل ذكي يعتمد على الفهم السياقي، واسترجاع المعرفة، واستخدام الأدوات، والتحقق من المصطلحات، ومراجعة الجودة .يوضح هذا المقال أن وكلاء الترجمة المدعومين بالنماذج اللغوية الكبيرة يمثلون نقلة نوعية في مجالات حساسة مثل الدبلوماسية، والعلوم، والهندسة، والقانون، والطب، والأعمال .وتزداد أهمية هذا التحول بسبب. عدم التوازن اللغوي في محتوى الويب، حيث تهيمن اللغة الإنكليزية على نسبة كبيرة من المواقع، بينما تمثل الفرنسية والعربية نسبا أقل بكثير، مما يؤثر في كمية المعرفة المتاحة للنماذج الذكية .لذلك، لا يكفي أن يكون وكيل الترجمة طليقا لغويا، بل يجب أن يكون موثوقا، ومؤسسا على معرفة متخصصة، وقادرا على طلب مراجعة بشرية عندما ترتفع المخاطر .كما يناقش المقال أهمية التقييم المستمر، سواء من خلال الإنسان في الحلقة، أم من خلال أدوات تقييم آلية مثل: (ProofAgent Harness)؛ لضمان جودة الترجمة، ودقتها، وسلامتها في السياقات عالية الحساسية.

الكلمات المفتاحية :وكلاء الذكاء الاصطناعي؛ الترجمة الآلية؛ النماذج اللغوية الكبيرة؛ الترجمة الوكيلة؛ العربية؛ الفرنسية؛ التحقق؛ الإنسان في الحلقة.

Introduction : Translation has always been more than linguistic substitution. A competent translator does not simply replace words; the translator interprets intent, preserves meaning, adapts register, resolves ambiguity, and protects the social, legal, or technical consequences of communication. In diplomatic communication, a misplaced modal verb can alter the perceived position of a state. In a scientific article, an imprecise term can distort a method or invalidate reproducibility. In technical documentation, an inconsistent component name can create operational risk. In legal and compliance contexts, translation may determine whether obligations, rights, and responsibilities are correctly understood across jurisdictions.

Traditional machine translation made this process faster and more scalable. Statistical machine translation and neural machine translation improved fluency and enabled large-scale multilingual communication. The Transformer architecture significantly accelerated this progress by replacing recurrence and convolution with attention-based sequence modeling, demonstrating strong results on machine translation tasks and becoming the foundation of modern large-scale language models [1]. However, most machine translation systems remained centered on a narrow task: given a source text, produce a target text. Such systems may be highly optimized, but they usually lack an explicit planning loop, external verification tools, dynamic retrieval, user-specific policy memory, and a structured mechanism for escalation.

The new generation of LLM-powered AI agents changes this picture. Large language models demonstrated that scaling and instruction-based generalization can expand the range of language tasks that a single model can perform [2]. In an agentic setting, however, the model is not simply a text generator. It becomes part of an operational system composed of memory, tools, retrieval, policies, and a control loop. In translation, this means the system can move from one-shot conversion to an intelligent workflow: understand the domain, identify critical terms, retrieve trusted references, apply a domain glossary, generate candidate translations, verify consistency, measure confidence, and request human review when necessary.

The importance of this shift becomes clearer when considering the linguistic structure of the web. According to W3Techs, as of June 13, 2026, English accounted for 49.7 percent of websites whose content language was known, French accounted for approximately 4.5 to 4.6 percent, and Arabic accounted for 0.6 percent [10]. This imbalance shapes training data, retrieval results, terminology coverage, cultural representation, and the knowledge available to translation systems. A translation agent operating between English, French, and Arabic must therefore do more than map sentences. It must actively compensate for uneven data availability, low-resource terminology gaps, dialect variation, and domain-specific knowledge scarcity.

This paper argues that AI agents will not merely improve translation; they will redefine the translation function. The translator of the future will increasingly be a hybrid system: a human expert supported by an AI agent that handles retrieval, terminology management, draft generation, consistency checking, and evaluation. In low-risk contexts, the agent may operate with high autonomy. In high-risk contexts, it must remain under human supervision and be subject to rigorous automated evaluation. This creates a new frontier for translation studies, computational linguistics, enterprise AI, and AI governance.

Related Work : The field of machine translation has progressed through several major phases. Rule-based systems attempted to encode linguistic knowledge manually. Statistical machine translation introduced probabilistic alignment across parallel corpora. Neural machine translation then modeled translation as a sequence-to-sequence learning problem, improving fluency and generalization. Transformer architectures further advanced this trajectory by enabling attention-based modeling at scale, with direct implications for translation quality and training efficiency [1]. Modern multilingual systems such as Meta's No Language Left Behind (NLLB) expanded this ambition to 200 languages, including many low-resource languages [3]. Google's work on building translation systems for the next thousand languages similarly emphasized web-mined datasets, language identification, multilingual modeling, and the limitations of evaluation in data-sparse conditions [4].

Large language models introduced a second shift. Unlike specialized translation engines, LLMs are trained on broad multilingual corpora and can perform translation as part of a broader reasoning and generation capability [2]. They can explain translation choices, adapt tone, follow style guides, preserve formatting, and handle multi-turn clarification. However, LLMs are also vulnerable to hallucination, overconfidence, cultural flattening, and inconsistent terminology. These limitations are amplified in specialized domains where factual precision matters.

Retrieval-Augmented Generation (RAG) addressed part of this problem by connecting language models to external knowledge sources. Instead of relying only on parametric memory, a RAG system can retrieve relevant documents at runtime and generate outputs grounded in those documents [5]. More recently, Agentic RAG has extended this idea by adding planning, tool use, memory, reflection, and multi-step retrieval strategies [9]. For translation, this transition is highly relevant because domain translation often requires more than bilingual mapping. It requires access to glossaries, standards, previous translations, regulations, scientific taxonomies, institutional style guides, and context-specific constraints.

The emerging literature on AI agents provides the final building block. ReAct showed that language models can interleave reasoning and acting, allowing them to use external information while maintaining a task trajectory [6]. Toolformer demonstrated that language models can learn to call external tools through simple APIs and incorporate tool outputs into future predictions [7]. Bousetouane's Agentic Systems framework further positions LLM agents as the cognitive backbone of vertical AI systems and proposes reusable building blocks for industry-specific agent design [8]. In translation, these capabilities enable workflows that resemble expert translation practice: terminology research, background reading, drafting, revision, quality assurance, and final approval.

Evaluation remains a central challenge. Translation quality has traditionally been assessed with automatic metrics and human review. These methods remain useful, but they are insufficient for agentic translation. An agent must be evaluated not only on final fluency but also on retrieval quality, evidence use, consistency across turns, correct tool use, policy compliance, cultural adequacy, and appropriate escalation. ProofAgent Harness framework is relevant here because it treats the agent as a system under behavioral pressure, using adversarial multi-turn evaluation, multi-juror scoring, and turn-level audit [12], [13].

AI Agents as the New Translator: The phrase "AI agent as translator" should not be misunderstood as a claim that all human translators will be replaced. The stronger and more accurate claim is that AI agents transform the translation function from a narrow linguistic service into a knowledge-mediated decision workflow. A conventional translation engine receives a source segment and returns a target segment. An agentic translator receives a goal, interprets constraints, gathers knowledge, manages uncertainty, and produces a translation artifact with evidence and governance logic.

A mature translation agent can operate through several stages. First, it classifies the task: diplomatic note, technical manual, scientific abstract, legal clause, medical instruction, marketing copy, or internal enterprise document. Second, it identifies the level of risk: low-risk informational translation, medium-risk professional translation, or high-risk translation with legal, safety, diplomatic, or reputational consequences. Third, it retrieves supporting context: prior translations, official terminology, standards, regulatory documents, and domain glossaries. Fourth, it generates a draft translation while preserving tone, register, and formatting. Fifth, it checks the draft against terminology rules, domain facts, and style constraints. Sixth, it produces a confidence estimate and flags passages requiring human review.

This workflow is closer to how expert translators work. Human translators do not translate in isolation; they search, compare, ask questions, consult domain experts, and revise. AI agents make this process programmable and scalable. They can perform repetitive checking, maintain consistency over long documents, align with institutional style, and detect uncertainty. In enterprise settings, they can also connect to translation memories, document management systems, policy repositories, and customer-specific knowledge bases.

The key advantage is grounding. A general LLM may translate a scientific expression fluently but incorrectly if it lacks domain context. A grounded translation agent can retrieve the relevant technical standard, compare equivalent terminology in previous documents, and verify the target term against trusted sources. For example, in a medical translation workflow, the agent may check drug names, contraindications, dosage units, and patient-facing readability. In an engineering workflow, it may verify part numbers, safety warnings, measurement units, and procedural sequence. In a diplomatic workflow, it may preserve intentional ambiguity, politeness strategies, and institutional language.

Another advantage is controllable autonomy. A translation agent can operate in different modes. In assistant mode, it supports a human translator with terminology suggestions and quality checks. In co-pilot mode, it drafts and revises under human supervision. In autonomous mode, it translates low-risk content with automated quality checks. In certified mode, it produces a translation only after passing evaluation gates and human approval. This layered autonomy is important because translation is not one task; it is a family of tasks with different risk profiles.

Web Language Imbalance and Its Impact on Translation Agents: The web is not linguistically neutral. Online content is heavily concentrated in a small number of languages, and this concentration influences both model training and retrieval. English remains the dominant language of web content. French has meaningful but much smaller representation. Arabic, despite being spoken by hundreds of millions of people across many countries, represents a very small share of measured web content. W3Techs reported English at 49.7 percent of known-content-language websites, French at approximately 4.5 to 4.6 percent, and Arabic at 0.6 percent in June 2026 [10]. This matters because AI systems learn patterns, terminology, facts, and cultural associations from available data.

The gap between English, French, and Arabic creates several consequences for translation agents. First, English-to-Arabic or Arabic-to-English translation may suffer from uneven terminology coverage. Scientific and technical terms may be abundant in English, moderately available in French, and fragmented or inconsistent in Arabic. Second, Arabic introduces dialectal diversity. Modern Standard Arabic is used in formal writing, but everyday communication varies widely across North Africa, the Gulf, Egypt, the Levant, and other regions. A translation agent must understand whether the target is formal Arabic, simplified Arabic, a regional dialect, or a hybrid communication style.

Third, retrieval quality differs across languages. A query in English may return many relevant sources, while the same query in Arabic may return fewer, older, or less authoritative documents. If a translation agent relies naively on web retrieval, it may over-ground translations in English sources and underrepresent Arabic conceptual frameworks. This creates a subtle form of linguistic dependency: the target language becomes a surface rendering of English knowledge rather than a fully grounded linguistic and cultural artifact.

Fourth, underrepresentation affects evaluation. If evaluation datasets are dominated by English-centric source sentences, translation systems may appear strong while failing on culturally specific, domain-specific, or regionally grounded content. Recent critiques of multilingual translation benchmarks show that evaluation data can remain culturally and domain biased, even when quality-control procedures are present [11]. For Arabic and other underrepresented languages, evaluation must include authentic domain content, local terminology, and human reviewers familiar with both language and subject matter.

Agentic translation can reduce these risks, but only if designed correctly. The agent should not simply retrieve more English content. It should retrieve multilingual sources, official terminology banks, domain corpora, institutional documents, and human-validated translation memories. It should track provenance: which source supported which term, which glossary governed which phrase, and where uncertainty remains. In low-resource contexts, the agent should explicitly report gaps rather than hiding them behind fluent language.

Domain Value: Diplomatic, Scientific, Technical, and Enterprise Translation: The value of agentic translation is strongest in domains where language carries operational consequences. Diplomatic translation is one example. Diplomatic language often relies on controlled ambiguity, protocol, politeness, and historical sensitivity. A literal translation may be accurate at the sentence level but harmful at the strategic level. An AI translation agent for diplomacy should retrieve prior communiques, align with official terminology, preserve intentional nuance, and flag phrases that may shift tone or position.

Scientific translation is another high-value domain. Researchers increasingly publish, collaborate, and communicate across languages. However, scientific translation requires accuracy in methods, variables, taxonomies, units, and claims. An agentic translator can check whether a term is used consistently across the literature, preserve citation context, and avoid converting tentative claims into stronger statements. It can also help Arabic and French scientific communities access English literature while producing higher-quality local-language scientific materials.

Technical and engineering translation requires strict consistency. Manuals, safety instructions, installation procedures, and maintenance documents often use controlled language. A translation agent can enforce a glossary, preserve warnings, validate units, and compare a translated procedure against the original sequence. This reduces the risk that fluent translation introduces operational errors. In industries such as energy, transportation, manufacturing, and aerospace, this capability has direct safety implications.

Legal and compliance translation requires both linguistic accuracy and jurisdictional awareness. A contract clause, privacy notice, financial disclosure, or regulatory policy cannot be translated as ordinary prose. It must preserve legal intent and align with jurisdiction-specific terminology. Here, the agent should retrieve applicable legal definitions, compare clauses with approved templates, and route high-risk language to a qualified human reviewer. It should also maintain an audit trail to show how terms were selected and which sources were used.

Enterprise translation brings these capabilities together at scale. Global organizations translate customer support content, HR policies, training materials, product pages, internal procedures, and technical documentation. A translation agent connected to enterprise knowledge can preserve brand voice, policy constraints, and product terminology. It can also learn from human corrections over time, improving consistency across departments and regions.

Reference Architecture for an Agentic Translation System: A practical agentic translation system can be organized around six layers. The first layer is the user interface, where a translator, diplomat, scientist, engineer, or business user submits the translation request and specifies the target audience, language variant, and risk level. The second layer is the orchestration layer, where the agent plans the translation workflow, chooses tools, and decides whether to retrieve documents, apply glossaries, or request human review. This orchestration layer reflects the broader design pattern of agentic systems: a cognitive core supported by tools, memory, external knowledge, and domain-specific skills [8].

The third layer is the knowledge layer. This includes domain glossaries, translation memories, approved bilingual corpora, style guides, legal templates, scientific taxonomies, technical standards, and prior institutional translations. The fourth layer is the tool layer. Tools may include terminology extraction, named entity recognition, document alignment, unit conversion, citation retrieval, grammar checking, readability scoring, and consistency comparison. The fifth layer is the generation layer, usually powered by an LLM or a specialized translation model. The sixth layer is the evaluation and governance layer, which checks the output before release.

This architecture is important because the translation output should not be treated as a single generative act. It should be treated as an artifact produced by a workflow. Each workflow step can be logged, inspected, and improved. For example, if a translation fails because the wrong glossary was retrieved, the problem is not only the model; it is the retrieval and orchestration logic. If a translation is fluent but legally unsafe, the problem may be insufficient risk classification or lack of escalation. Agentic translation therefore requires system-level engineering, not just model selection.

A simplified workflow would look as follows: intake the request; classify domain and risk; retrieve trusted references; extract critical terms; generate a first draft; check terminology and facts; produce a revised translation; score the result; escalate uncertain passages; and store approved outputs in translation memory. This creates a closed loop in which every validated translation improves the future system.

The architecture also supports specialization. A diplomatic translation agent may have a policy memory and official-state terminology database. A scientific translation agent may connect to scholarly databases and domain ontologies. A healthcare translation agent may enforce patient-safety constraints and plain-language guidelines. A technical translation agent may connect to product databases and safety standards. The core agentic pattern remains the same, but the knowledge base, tools, evaluation criteria, and escalation thresholds change by domain.

Evaluation, Human-in-the-Loop, and ProofAgent-Style Auditing: The strongest translation agent is not the one that sounds most fluent. It is the one that can be trusted under domain pressure. Translation failures are often subtle: a missing negation, a softened obligation, a changed degree of certainty, an inconsistent term, a wrong unit, or a culturally inappropriate tone. These failures may not be detected by surface-level fluency metrics. They require structured evaluation.

Human-in-the-loop evaluation remains necessary, especially in high-risk domains. Human reviewers bring cultural judgment, domain expertise, institutional memory, and accountability. However, human review alone does not scale. Organizations need automated tools that can stress-test translation agents, replay failures, compare versions, and detect regressions. This is where agent evaluation frameworks become part of the translation infrastructure.

A ProofAgent-style evaluation approach can treat the translation agent as a system under test. Instead of scoring one translated sentence, the evaluator can run multi-turn scenarios: ambiguous diplomatic language, conflicting glossary entries, missing context, adversarial user instructions, domain-specific terminology traps, and policy constraints. The ProofAgent Harness framework is designed for adversarial, multi-turn AI agent evaluation and introduces Adversarial Multi-Juror Scoring with Turn-Level Audit to connect judgments to behavioral evidence [12], [13].

For translation, the evaluation rubric should include at least ten dimensions. First, semantic fidelity: does the target text preserve the meaning of the source? Second, terminology accuracy: are domain terms translated according to approved glossaries? Third, register and tone: does the translation match diplomatic, scientific, legal, technical, or public-facing requirements? Fourth, cultural adequacy: does the translation avoid inappropriate literalism or cultural distortion? Fifth, factual grounding: are factual claims supported by trusted references? Sixth, consistency: are terms and entities handled consistently across the document? Seventh, uncertainty awareness: does the agent flag unclear or risky segments? Eighth, tool use: did the agent retrieve and apply the right sources? Ninth, policy compliance: did it respect confidentiality, privacy, and domain constraints? Tenth, escalation: did it route high-risk cases to human review?

Evaluation should also include adversarial tests. A user may ask the translation agent to ignore an official glossary, simplify a legal obligation in a misleading way, remove a safety warning, or translate confidential text through an unauthorized tool. A robust agent should resist these instructions. It should preserve policy constraints and explain why certain requests cannot be fulfilled. This is not merely a translation problem; it is an AI governance problem.

In enterprise deployment, evaluation metrics should include task success, relevance, drift and memory stability, hallucination and factuality, instruction following, safety, policy compliance, manipulation susceptibility, and coordination quality. These metrics can be adapted directly to translation agents, especially when translation becomes a multi-step workflow involving retrieval, tools, memory, and human approval.

Discussion: From Translator Replacement to Translation Intelligence: The most productive way to think about AI agents in translation is not replacement, but translation intelligence. Translation intelligence includes the ability to understand context, manage domain knowledge, reason over uncertainty, preserve institutional meaning, and produce an auditable communication artifact. In this sense, agentic translation expands the translator's role rather than simply automating it.

Human translators may increasingly become supervisors of translation systems, curators of domain knowledge, designers of glossaries, reviewers of high-risk passages, and evaluators of agent behavior. This shift resembles what is happening in other professional domains: the human moves from performing every micro-task manually to designing, supervising, and validating intelligent workflows. The value of human expertise does not disappear; it moves upstream and becomes more strategic.

This transition also creates new responsibilities. Organizations deploying translation agents must define what autonomy is allowed, which domains require human approval, which data sources are trusted, how confidential documents are handled, and how failures are reported. They must also monitor bias across languages. If English dominates training and retrieval, translation agents may unconsciously privilege English conceptual structures. For Arabic, this may result in translations that are grammatically correct but culturally or institutionally weak. For French, it may affect legal or administrative nuance. The solution is not only better models; it is better grounding, better evaluation, and better human oversight.

The future translation function will likely include three layers. The first layer is automated translation for low-risk content. The second layer is agent-assisted professional translation for domain documents. The third layer is certified translation intelligence for high-risk contexts, where every output is grounded, evaluated, logged, and approved. This layered model respects both the power of AI and the complexity of language.

Conclusion: AI agents represent a major turning point in translation. The field is moving beyond sentence-level machine translation toward autonomous, domain-grounded, tool-using translation workflows. This shift is driven by LLMs, retrieval, memory, tool use, and agent orchestration. It is also driven by the real-world limitations of the current web: English dominates online content, while languages such as French and Arabic have much smaller shares, creating imbalances in terminology, knowledge availability, and evaluation coverage [10].

The opportunity is significant. AI translation agents can support diplomacy, science, technology, law, healthcare, and enterprise communication by grounding translations in trusted knowledge, maintaining consistency, and scaling multilingual access. But the risk is equally significant. Fluent translation can hide factual, cultural, legal, and safety failures. Therefore, agentic translation must be paired with human-in-the-loop review and automated evaluation frameworks such as ProofAgent Harness [12], [13].

The central claim of this paper is that translation is becoming an intelligent, auditable, and domain-aware function. The translator of the future is not only a person or a model. It is a human-agent system: a collaborative structure where human judgment, domain knowledge, AI generation, tool-based verification, and continuous evaluation work together. In this new paradigm, the goal is not only to translate words, but to preserve meaning, trust, and responsibility across languages.

References:

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems. https://arxiv.org/abs/1706.03762

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems. https://arxiv.org/abs/2005.14165

[3] Meta AI. (2022). No Language Left Behind: Open-source models for translation across 200 languages. https://ai.meta.com/research/no-language-left-behind/

[4] Bapna, A., Caswell, I., Kreutzer, J., Firat, O., van Esch, D., Siddhant, A., Niu, M., Baljekar, P., et al. (2022). Building Machine Translation Systems for the Next Thousand Languages. arXiv:2205.03983. https://arxiv.org/abs/2205.03983

[5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. https://arxiv.org/abs/2005.11401

[6] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629. https://arxiv.org/abs/2210.03629

[7] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., and Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. Advances in Neural Information Processing Systems. https://arxiv.org/abs/2302.04761

[8] Bousetouane, F. (2025). Agentic Systems: A Guide to Transforming Industries with Vertical AI Agents. arXiv:2501.00881. https://arxiv.org/abs/2501.00881

[9] Singh, A., Ehtesham, A., Kumar, S., Khoei, T. T., and Vasilakos, A. V. (2025). Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. arXiv:2501.09136. https://arxiv.org/abs/2501.09136

[10] W3Techs. (2026). Usage statistics of content languages for websites. Accessed June 13, 2026. https://w3techs.com/technologies/overview/content_language

[11] Taguchi, C., Mai, S., Kurabe, K., Sakai, Y., Agyei, G., Eslami, S., and Chiang, D. (2025). Languages Still Left Behind: Toward a Better Multilingual Machine Translation Benchmark. arXiv:2508.20511. https://arxiv.org/abs/2508.20511

[12] Bousetouane, F. (2026). ProofAgent Harness: Open Infrastructure for Adversarial Evaluation of AI Agents. arXiv:2605.24134. https://arxiv.org/abs/2605.24134

[13] ProofAgent. (2026). ProofAgent Harness: Open-source infrastructure for adversarial AI agent testing. https://www.proofagent.ai/harness