Short: The Monolingual Assumption: What Web Agents Pay to Work Beyond English

Hannah B. Pasandi
University of California, Berkeley
Berkeley, CA, USA
h.pasandi@berkeley.edu

Haniyeh B. Pasandi
University of Maryland, College Park
College Park, MD, USA
hpasandi@umd.edu

Abstract

Some languages pay more tokens than English for identical content. For model APIs, this token premium is well established. Inside a web agent, it is unmeasured. There, the tokenizer's bill is only the entry fee. Context windows truncate sooner. Extractors tuned on English layouts fail in new ways. Retries multiply the base cost. And actions grounded in translated text can miss their targets on the source page. We introduce a methodology that isolates language as the only variable. Many sites publish the same page in several languages. These parallel editions hold content and stack constant, so each pair becomes a controlled experiment. As a first step, we complete a 23-language pre-study on one parallel document. It sets the floor of the tax. Under the English-centric BPE of the first agent era, identical content costs 2–23× more tokens outside English. A modern 256k multilingual vocabulary narrows the gap but does not close it: Farsi still pays 1.4×, Hindi 2.0×, and Burmese 6.2×. The tax follows vocabulary allocation, not script. Chinese reaches English parity while alphabetic Georgian pays 3.9×. An equitably allocated research tokenizer flattens the curve to at most 2.2×. The tax, in short, is a choice. We release TwinPages, a corpus builder over the top 1,000 ranked origins and their language editions; an agent-cost protocol over it; and a four-clause language contract for agent runtimes, whose final clause checks every non-English extraction against the parallel English page.

Keywords: web agents, multilingual, tokenization, ML for systems, measurement, deployment

1 Introduction

Ask a frontier assistant a question in Farsi and the answer often reads like English wearing a Farsi costume: translated phrasing, not native prose. Speak the question instead of typing it and the outcome can be starker, transcribed correctly by one assistant and not recognized at all by another. These surface symptoms have a stack-level cause. Every layer a web agent runs on, the tokenizer's vocabulary, the speech front-end, the extractor's readability heuristics, the context budget, the retry policy, and the invoice, was tuned on English, and none of them declares it. We call this the *monolingual assumption*: language is treated as a property of the model when it is in fact a property of the stack.

The model-side half of the story is known. Tokenizer studies established that commercial vocabularies charge some languages several times more tokens for the same content [1, 26, 30]. But those studies price chat: text in, text out. Web agents [10, 37, 40] add a stack on top of the tokenizer, and every layer of it inherits the multiplier. A context window holds proportionally fewer pages [17]. Truncation, and the retries it triggers, arrive proportionally sooner. Extraction pipelines tuned on Western layouts [5] fail on scripts they were never tested on, and an agent that reasons over a translation acts on text that no longer exists in the source DOM. None of this compounding has been measured, because measuring it requires holding the page constant while varying only its language.

The web supplies exactly that instrument. Sites with official parallel versions, encyclopedias, government portals, global marketplaces, publish the same page in many languages. Such pairs hold content and stack constant while language varies, the way a controlled study isolates a single factor (Figure 1), and the population they are drawn from can be made representative by starting from the most accurate ranked list of the web [12, 31].

Contributions. This paper makes four.

- **A methodology.** TwinPages treats parallel-language page pairs as controlled experiments, drawn from the top 1,000 ranked origins and their official language editions, with a released manifest builder (§2).
- **A measured floor.** A completed 23-language pre-study under three production-grade tokenizers quantifies the tax on identical content and yields four findings, including that the tax tracks vocabulary allocation, not script, and is therefore a design choice (§3).
- **A separation protocol.** An agent-cost protocol over TwinPages whose paired design separates the stack's own contribution, extraction failures, retries, truncation, from the tokenizer floor (§4).
- **A mitigation contract.** Four clauses an agent runtime can enforce today, ending with a differential check that turns the measurement instrument itself into a runtime safeguard (§5).

Broader impact. The population that pays this tax is the population the next billion users come from. The languages at the expensive end of Figure 2 are spoken predominantly in markets already underserved by language technology [15],

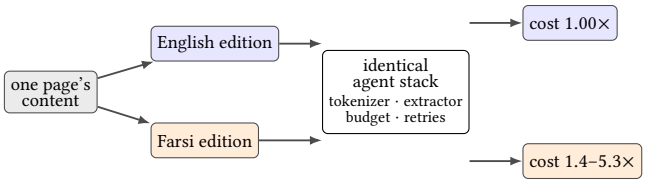

Figure 1. The methodology in one picture. Parallel-language editions of the same page hold content and stack constant, so the price difference is attributable to language alone (Farsi multipliers measured, §3).

and the premium is regressive three times over: the same information costs more per token [1], fits less of it in every context window, and compounds into the energy and carbon of test-time compute [25]. Because agents act as well as read, the assumption also converts cost inequity into capability inequity, a booking flow that an English-speaking user's agent completes may fail for a Farsi speaker's agent on the same site. Making the tax measurable is the precondition for billing it fairly, budgeting for it explicitly, and engineering it away; the contract in §5 is a first concrete step, and the corpus makes progress auditable.

2 TwinPages: Same Page, Different Price

Ranked population. The site population comes from the Chrome UX Report top lists, shown by direct comparison against CDN-side ground truth to be the most accurate ranking of popular websites [31]; our snapshot [12] yields a rank-1000 bucket of exactly 1,000 origins, and per-country editions of the same list provide language-market strata. **Page pairs.** A pair is two URLs with officially parallel content: Wikipedia language links, institutional portals with language switchers, and any ranked origin exposing `hreflang` alternates. The released builder assembles a manifest of such pairs per language. **Tokenizer floor.** Before live pages, we measure the floor on a document that exists in parallel across hundreds of languages, the Universal Declaration of Human Rights [34], tokenized under three vocabularies chosen to span the design space: the English-centric GPT-2 BPE that defined the first agent era [29, 33], the equitably allocated research tokenizer XLM-R [9], and a modern 256k-vocabulary production SentencePiece [13, 16]. Code, data, and the manifest builder are released; every number in §3 reproduces with one script.

3 Pre-Study Findings

Figure 2 shows all 23 languages; Table 1 gives exact values for a representative subset.

F1. Under the English-centric BPE of the first agent era, identical content costs 2–23× more tokens outside English; the median non-English language pays roughly 5×.

F2. A modern 256k multilingual vocabulary narrows but does not close the gap: 1.2–6.2×, and the poorest-resourced scripts pay the most.

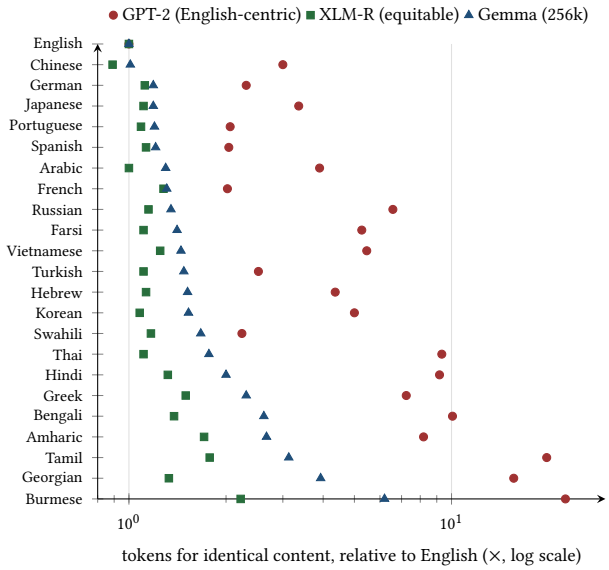

Figure 2. Identical content, 23 languages, three tokenizers. The English-centric vocabulary charges up to 23×; a 256k multilingual vocabulary still charges up to 6.2×; an equitably allocated one flattens the curve to ≤2.2×.

Table 1. Exact multipliers for selected languages (full 23-language CSV released).

Language	GPT-2 ×	XLM-R ×	Gemma ×
Chinese	3.00	0.89	1.01
Arabic	3.90	1.00	1.30
Farsi	5.27	1.11	1.41
Hindi	9.18	1.32	2.00
Tamil	19.73	1.78	3.13
Georgian	15.59	1.33	3.93
Burmese	22.55	2.22	6.20

F3. The tax is allocation, not script. Chinese reaches English parity (1.01×) while alphabetic Georgian pays 3.9×: cost tracks vocabulary investment, which tracks training-data economics [30, 39].

F4. The tax is a choice. A tokenizer built for equitable coverage flattens the curve to at most 2.2× across all 23 languages, and token-free architectures eliminate it by construction at the cost of longer sequences [8, 35].

What the floor buys downstream. Every stack budget inherits the multiplier. Measured against our parallel document under the 256k vocabulary, a 128k-token context holds 62 copies in English, 44 in Farsi, 31 in Hindi, and 10 in Burmese, so truncation, and the retry loops it triggers, arrive up to 6× sooner for the same information. At per-token prices the same content bills a 41% premium in Farsi and a 520% premium in Burmese before any agent inefficiency is counted, and the premium compounds into the energy and carbon of test-time compute [25]. Findings F1–F4 are the floor; the protocol below measures how far live agents rise above it.

4 Agent-Cost Protocol

[Planned; harness released with the corpus builder.] Three task classes run identically on both pages of every pair: extract five key facts as JSON; locate the section answering a fixed question; and complete a three-step interactive flow where the origin offers one. Per run the harness logs input and output tokens, model calls, retries, truncation events, extraction-schema validity, and rubric-scored success, at temperature zero with three repetitions and the task text held in English while the page stays in situ. The reported unit is the per-pair language multiplier (median and IQR per language), with a failure taxonomy of extraction garbage, truncation, translate-then-act mismatch, and translationese output, following the practice of building failure taxonomies before mechanisms [7]. Because every run is paired against the English edition, agent-layer inefficiency separates cleanly from the tokenizer floor of §3: the floor predicts the multiplier if the stack were language-neutral, and the residual above it is the stack's own contribution.

5 The Language Contract

The pre-study already implies four clauses an agent runtime can enforce today (Figure 3). **C1, script-aware budgeting:** context and cost budgets scale by the measured per-language multiplier instead of a universal token constant, so a Farsi task is not silently granted 71% of an English task's effective window. **C2, extraction with refusal:** when extractor confidence collapses on a script it was not tuned for, the runtime refuses and re-fetches raw content rather than hallucinating structure, converting silent garbage into a loggable event. **C3, act-on-source binding:** reasoning may use a translation, but every action grounds in a source-DOM span, eliminating the translate-then-act mismatch by construction. **C4, differential checking:** where a parallel English edition exists, the non-English extraction is validated against it, and divergence beyond a threshold triggers C2's refusal path. C4 turns the paper's measurement instrument into a runtime safeguard, and it is the clause the TwinPages manifest makes deployable.

6 Discussion

The speech boundary. The assumption does not stop at text. Whether spoken Farsi becomes text at all depends on which speech front-end a product ships: massively multilingual ASR trained on web-scale weak supervision covers the language well [27, 28], while stacks descended from English-centric benchmarks [22] or thin platform dictation may not recognize it, and community corpora document exactly this coverage gap [4]. In our own use, the same spoken Farsi query is transcribed by one frontier assistant and unrecognized by another, which means the tax begins before the first token is billed. The protocol extends naturally: twin utterances, the same request spoken in two languages, measure the speech layer the way twin pages measure the text stack, and emerging audio-language-model evaluations provide the fairness axes to score against [3].

Assistant-level variation. Users report, and we observe, that frontier assistants differ visibly in non-English fluency: one produces native-register Farsi while another produces translated-sounding prose for the same prompt. This is unsurprising once language is seen as a stack property, since assistants differ in tokenizer allocation (measured here across three vocabularies), in multilingual training mixture [6, 13, 21], and in speech integration. We deliberately make no head-to-head product claims in this paper; the twin design turns the question into a measurable one, same prompt, two languages, scored for translationese and task success across assistants, and we leave that product study to the harness rather than to anecdote.

Beyond tokens. Three further stack layers carry the assumption and are visible in our failure taxonomy but not yet priced: right-to-left layout and mixed-direction DOMs that confuse span binding; locale infrastructure, date, number, and currency formats plus geo-redirects that change what page a language user even receives; and code-switched content, common in real usage, that no single-language multiplier describes. Each is a candidate fifth clause once measured.

Limitations. The floor is measured on one parallel document whose register is formal; web registers will shift absolute multipliers, though F3's allocation result is register-independent by construction. Tokenizer cost is not end-to-end quality, cheap tokens do not guarantee good answers, which is exactly why the protocol scores task success alongside cost. Parallel editions are imperfect twins, since localization can add or drop content; the manifest screens pairs by structural similarity and C4 exists precisely because divergence is detectable. And the agent-layer numbers are the protocol's output, not this paper's; we mark them as such throughout.

7 Related Work

Token premiums and tokenizer design. Subword tokenization entered the modern stack through byte-pair encoding [33] and language-independent segmentation [16], and its inequity is now established at the model boundary: multilingual tokenizers charge unequal token counts for equal content [26], the inequality prices real commercial APIs at up to an order of magnitude [1], and segmentation quality predicts monolingual task performance [30]. The design space that creates and can close the gap is also mapped: vocabulary capacity can be allocated across languages deliberately [39], and token-free models remove the vocabulary entirely, trading the premium for longer byte or character sequences [8, 35]. Our floor study sits on this line and adds the systems consequence: the premium propagates into every

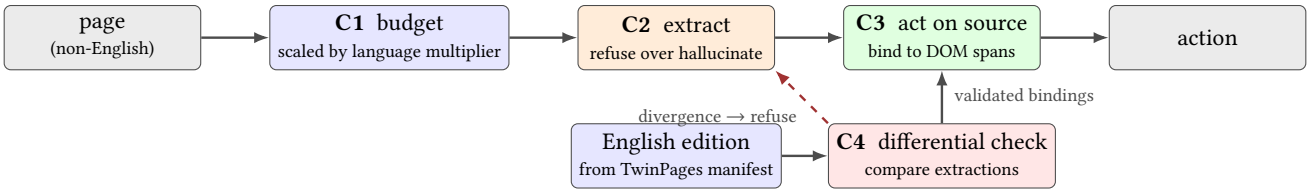

Figure 3. The language contract as a runtime pipeline. Budgets scale by measured multipliers (C1), extraction refuses rather than hallucinates (C2), actions bind to source text (C3), and the parallel English edition audits the result (C4), feeding refusal back into C2 on divergence.

budget of an agent stack. **Multilingual models and evaluation.** The multilingual model lineage runs from mBERT [11] through XLM-R [9], mT5 [36], mBART [19], and massively multilingual translation [20], with evaluation frameworks measuring cross-lingual generalization [14] and, more recently, generative quality across languages [2]. The field's own accounting of who is served shows the long tail systematically excluded [15]. These efforts evaluate models; we evaluate the deployed stack around them, which none of these benchmarks touches. **Web agents and extraction.** Agent benchmarks establish realistic browsing and acting, WebShop for grounded commerce [37], Mind2Web for generalist tasks [10], WebArena for full environments [40], on top of the reasoning-and-tools substrate of ReAct and Toolformer [32, 38], with serving systems managing the context economics underneath [17, 18] and multi-agent failure studies supplying the taxonomy discipline we adopt [7]. These environments are overwhelmingly English; TwinPages is deliberately a lens over the existing web rather than a new environment, because the phenomenon under study is the deployed stack, including the extraction layer itself [5]. **Speech and audio.** The ASR lineage the assumption runs through begins at English audiobook benchmarks [22], is challenged by community multilingual corpora [4], and is partially answered by weakly supervised and thousand-language models [27, 28], while audio-language-model evaluation now includes fairness explicitly [3]. Our speech-boundary discussion imports exactly this axis into the agent stack. **Measurement and the adoption line.** The ranked population rests on the finding that CrUX outranks prior top lists against CDN ground truth [12, 31]. And this paper extends an adoption-obstacle program: companion work makes the plan the unit of agent accounting [23], writes an admission contract for learned systems on commodity hardware [24], and prices the carbon of test-time compute [25]; the language contract is the same move at the language boundary.

8 Conclusion

The monolingual assumption is not a model bug; it is a stack property, priced here at 1.2–6.2× even under a modern multilingual vocabulary and up to 23× under the vocabulary that trained the first agent era. Parallel pages make the price measurable on the live web, the protocol separates the stack's contribution from the tokenizer floor, and the four-clause contract gives runtimes a way to stop charging it silently. The next result this paper commits to is the agent-layer multiplier table over TwinPages, and the claim it will test is simple: beyond English, today's web agents pay the floor several times over.

References

[1] Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP '23)*. Association for Computational Linguistics, 9904–9923.

[2] Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. arXiv preprint arXiv:2303.12528.

[3] Ranya Aloufi, Srishti Gupta, Soumya Shaw, Battista Biggio, and Lea Schönherr. 2026. Evaluation of Audio Language Models for Fairness, Safety, and Security. arXiv preprint arXiv:2603.13262.

[4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC '20)*.

[5] Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL '21): System Demonstrations*.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS '20)*.

[7] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why Do Multi-Agent LLM Systems Fail? arXiv preprint arXiv:2503.13657.

[8] Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. CANINE: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics* 10 (2022).

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. 8440–8451.

[10] 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 (NeurIPS '23), Datasets and Benchmarks Track*.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '19)*. 4171–4186.

[12] Zakir Durumeric. 2026. crux-top-lists: Downloadable Snapshots of the Chrome Top Million Websites. GitHub repository, https://github.com/zakird/crux-top-lists, accessed July 2026.

[13] Gemma Team. 2024. Gemma: Open Models Based on Gemini Research and Technology. Google DeepMind technical report.

[14] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*.

[15] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL '20)*. 6282–6293.

[16] Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP '18): System Demonstrations*. 66–71.

[17] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP '23)*. 611–626. doi:10.1145/3600006.3613165

[18] Chaofan Lin, Zhenhua Han, Chengruidong Zhang, Yuqing Yang, Fan Yang, Chen Chen, and Lili Qiu. 2024. Parrot: Efficient Serving of LLM-based Applications with Semantic Variable. In *Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI '24)*. USENIX Association, 929–945.

[19] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 8 (2020).

[20] NLLB Team. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv preprint arXiv:2207.04672.

[21] OpenAI. 2023. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '15)*. IEEE, 5206–5210.

[23] Hannah B. Pasandi, Negar Arabzadeh, and Melissa Pan. 2026. PlanLayer: Making the Plan the Unit of Accounting for Agentic AI. Under submission, 5th Workshop on Practical Adoption Challenges of ML for Systems (PACMI '26).

[24] Hannah B. Pasandi, Mohammad Hosseini, and Sina Darabi. 2026. Learning on the GPUs the World Owns: A Fleet Contract for Learned Systems on Commodity Hardware. Under submission, 5th Workshop on Practical Adoption Challenges of ML for Systems (PACMI '26).

[25] Hananeh B. Pasandi and Tamer Nadeem. 2026. The Reasoning Tax: Profiling Test-Time Compute Carbon in Agentic AI Workloads. *ACM SIGENERGY Energy Informatics Review* 6, 2 (June 2026).

[26] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. In *Advances in Neural Information Processing Systems (NeurIPS '23)*. 36963–36990.

[27] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling Speech Technology to 1,000+ Languages. arXiv preprint arXiv:2305.13516.

[28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. OpenAI technical report.

[30] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL '21)*. 3118–3135.

[31] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. 2022. Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists. In *Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22)*. Association for Computing Machinery, 374–387. doi:10.1145/3517745.3561444

[32] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS '23)*.

[33] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*. 1715–1725.

[34] United Nations. 2026. The Universal Declaration of Human Rights: Translations Dataset. Office of the High Commissioner for Human Rights; distributed via the NLTK udhr2 corpus.

[35] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics* 10 (2022).

[36] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '21)*. 483–498.

[37] 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 (NeurIPS '22)*.

[38] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR '23)*.

[39] Bo Zheng, Li Dong, Shaohan Huang, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Allocating Large Vocabulary Capacity for Cross-lingual Language Model Pre-training. arXiv preprint arXiv:2109.07306.

[40] 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 *Proceedings of the 12th International Conference on Learning Representations (ICLR '24)*.