The Token Tax: Systematic Bias in Multilingual Tokenization

Jessica M. Lundin[1], Ada Zhang[2], Nihal Karim[2], Hamza Louzan[2], Victor Wei[2], David Ifeoluwa Adelani[3], Cody Carroll[2,4]

[1]Institute for Disease Modeling, Gates Foundation [2]Data Institute, University of San Francisco

[3]Mila – Quebec AI Institute, McGill University, and Canada CIFAR AI Chair [4]Dept. of Mathematics and Statistics, University of San Francisco

jessica.lundin@gatesfoundation.org — EACL 2026 Workshop Poster

The Problem: The Token Tax

Tokenizers trained predominantly on English fragment morphologically rich languages into more tokens, inflating compute costs and degrading accuracy. The transformer's quadratic $O(n^2)$ scaling turns this into a **"token tax"** — a prohibitive surcharge on training, inference, and CO_2 paid by billions of speakers.

30 pp
African langs trail English on average

4×
Training cost per 2× fertility increase

20-50%
Accuracy variance explained by fertility

18 pp
MMLU accuracy drop per extra token per word

Setup

- **Benchmark:** AfriMMLU — 16 African languages, 5 subjects, 9,000 multiple-choice questions
- **Models:** 10 LLMs including reasoning models (DeepSeek R1, o1) and general LLMs (Llama 3.1 405B, GPT-4o, Gemini 1.5 Pro, Claude Sonnet 3.5, and others)
- **Metric:** Fertility $F = T/W$ (tokens per word). Higher fertility $\Rightarrow$ worse performance and higher cost

Economic Impact of Token Inflation

Because transformer training scales quadratically with sequence length, a 2× increase in fertility produces a 4× increase in training time and cost.

Model	English	2× Fertility	5× Fertility
Llama 2 70B	\$5M	\$20M	\$125M
Llama 3 70B	\$24M	\$96M	\$600M
Llama 3.1 405B	\$105M	\$420M	\$2.6B

Training costs scale quadratically with fertility.

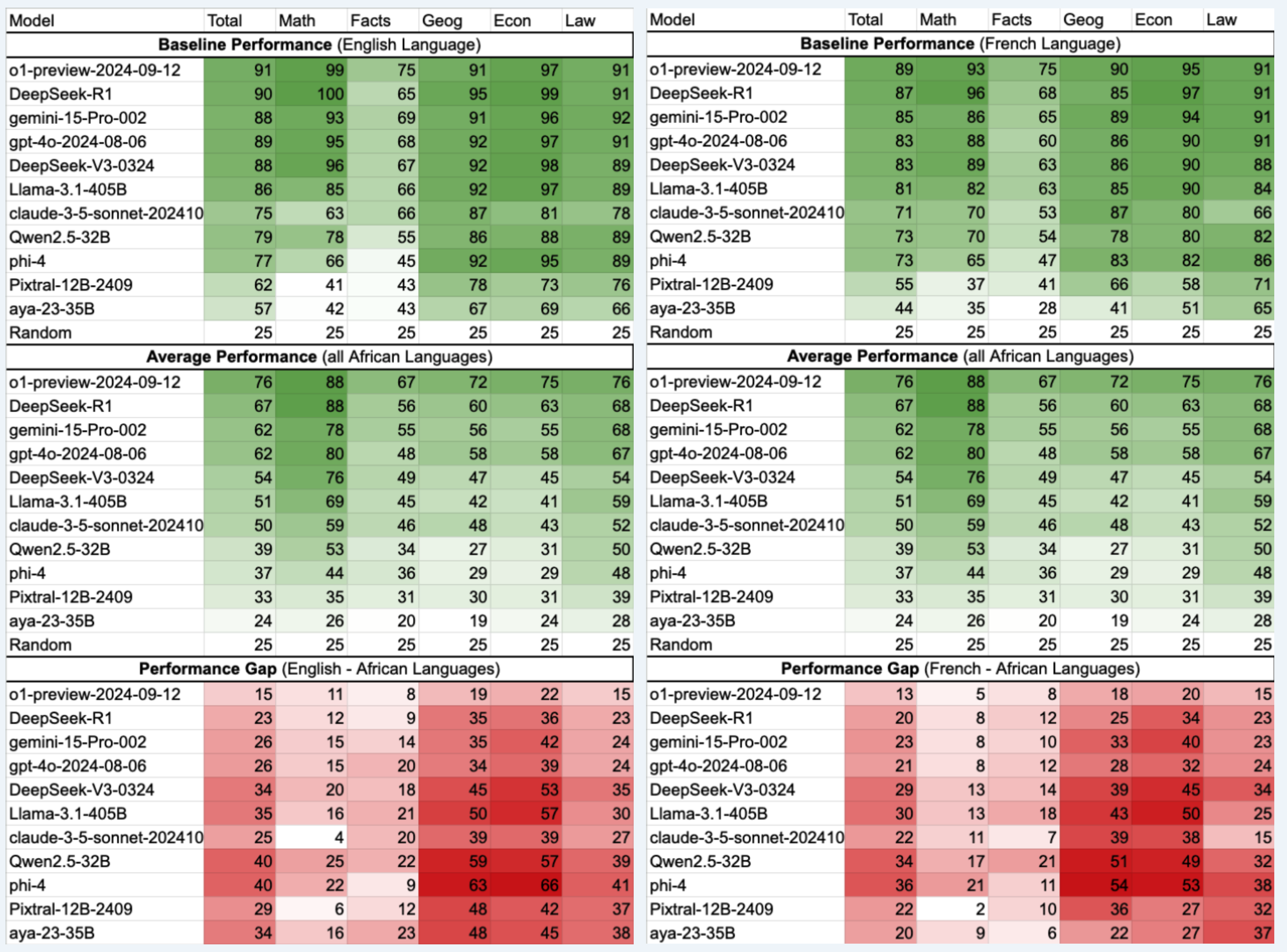
Provider	Model	English \$	Language X (∼2×)
OpenAI	GPT-4o	5 / 20	10 / 40
OpenAI	o4-mini*	4 / 16	8 / 32
Google	Gemini 2.5 Flash	0.30/2.50	0.60/5.00
Google	Gemini 2.5 Pro*	1.25/10	2.50/20
Anthropic	Claude 4 Sonnet	3 / 15	6 / 30
Anthropic	Claude 4 Opus*	15 / 75	30 / 150

Inference cost per 1M English-equivalent tokens (USD, input/output). *Reasoning models.

Key Takeaways

- Token fertility reliably predicts accuracy across all 10 models and 5 subjects
- Reasoning models (DeepSeek R1, o1) narrow but do not close the gap, improving African language performance by 8–12 points on average
- Doubling fertility quadruples training costs, creating a "token tax" that turns linguistic diversity into computational liability
- Addressing these inequities requires morphologically-aware tokenization, fair pricing, and expanded multilingual evaluation infrastructure

Performance Gaps Across Languages



Model	Total	Math	Facts	Geog	Econ	Law
Baseline Performance (English Language)						
o1-preview-2024-09-12	91	99	75	91	97	91
DeepSeek-R1	90	100	65	95	99	91
gemini-15-Pro-002	88	93	69	91	96	92
gpt-4o-2024-08-06	89	95	68	92	97	91
DeepSeek-V3-0324	88	96	67	92	98	89
Llama-3.1-405B	86	85	66	92	97	89
claude-3-5-sonnet-202410	75	63	66	87	81	78
Qwen2.5-32B	79	78	55	86	88	89
phi-4	77	66	45	92	95	89
Pixtral-12B-2409	62	41	43	78	73	76
aya-23-35B	57	42	43	67	69	66
Random	25	25	25	25	25	25
Average Performance (all African Languages)						
o1-preview-2024-09-12	76	88	67	72	75	76
DeepSeek-R1	67	88	56	60	63	68
gemini-15-Pro-002	62	78	55	56	55	68
gpt-4o-2024-08-06	62	80	48	58	58	67
DeepSeek-V3-0324	54	76	49	47	45	54
Llama-3.1-405B	51	69	45	42	41	59
claude-3-5-sonnet-202410	50	59	46	48	43	52
Qwen2.5-32B	39	53	34	27	31	50
phi-4	37	44	36	29	29	48
Pixtral-12B-2409	33	35	31	30	31	39
aya-23-35B	24	26	20	19	24	28
Random	25	25	25	25	25	25
Performance Gap (English - African Languages)						
o1-preview-2024-09-12	15	11	8	19	22	15
DeepSeek-R1	23	12	9	35	36	23
gemini-15-Pro-002	26	15	14	35	42	24
gpt-4o-2024-08-06	26	15	20	34	39	24
DeepSeek-V3-0324	34	20	18	45	53	35
Llama-3.1-405B	35	16	21	50	57	30
claude-3-5-sonnet-202410	25	4	20	39	39	27
Qwen2.5-32B	40	25	22	59	57	39
phi-4	40	22	9	63	66	41
Pixtral-12B-2409	29	6	12	48	42	37
aya-23-35B	34	16	23	48	45	38

(a) vs. English baseline

Model	Total	Math	Facts	Geog	Econ	Law
Baseline Performance (French Language)						
o1-preview-2024-09-12	89	93	75	90	95	91
DeepSeek-R1	87	96	68	85	97	91
gemini-15-Pro-002	85	86	65	89	94	91
gpt-4o-2024-08-06	83	88	60	86	90	91
DeepSeek-V3-0324	83	89	63	86	90	88
Llama-3.1-405B	81	82	63	85	90	84
claude-3-5-sonnet-202410	71	70	53	87	80	66
Qwen2.5-32B	73	70	54	78	80	82
phi-4	73	65	47	83	82	86
Pixtral-12B-2409	55	37	41	66	58	71
aya-23-35B	44	35	28	41	51	65
Random	25	25	25	25	25	25
Average Performance (all African Languages)						
o1-preview-2024-09-12	76	88	67	72	75	76
DeepSeek-R1	67	88	56	60	63	68
gemini-15-Pro-002	62	78	55	56	55	68
gpt-4o-2024-08-06	62	80	48	58	58	67
DeepSeek-V3-0324	54	76	49	47	45	54
Llama-3.1-405B	51	69	45	42	41	59
claude-3-5-sonnet-202410	50	59	46	48	43	52
Qwen2.5-32B	39	53	34	27	31	50
phi-4	37	44	36	29	29	48
Pixtral-12B-2409	33	35	31	30	31	39
aya-23-35B	24	26	20	19	24	28
Random	25	25	25	25	25	25
Performance Gap (French - African Languages)						
o1-preview-2024-09-12	13	5	8	18	20	15
DeepSeek-R1	20	8	12	25	34	23
gemini-15-Pro-002	23	8	10	33	40	23
gpt-4o-2024-08-06	21	8	12	28	32	24
DeepSeek-V3-0324	29	13	14	39	45	34
Llama-3.1-405B	30	13	18	43	50	25
claude-3-5-sonnet-202410	22	11	7	39	38	15
Qwen2.5-32B	34	17	21	51	49	32
phi-4	36	21	11	54	53	38
Pixtral-12B-2409	22	2	10	36	27	32
aya-23-35B	20	9	6	22	27	37

(b) vs. French baseline

Figure 1. Baseline performance shows English (a) and French (b) accuracy (in percentage points). The mean accuracy across all 16 African languages is shown in the middle charts of (a) and (b). The bottom charts of (a) and (b) show performance gaps between the African languages and higher-resource languages, though reasoning-oriented models narrow this gap.

Fertility Predicts Accuracy (Llama 3.1 405B)

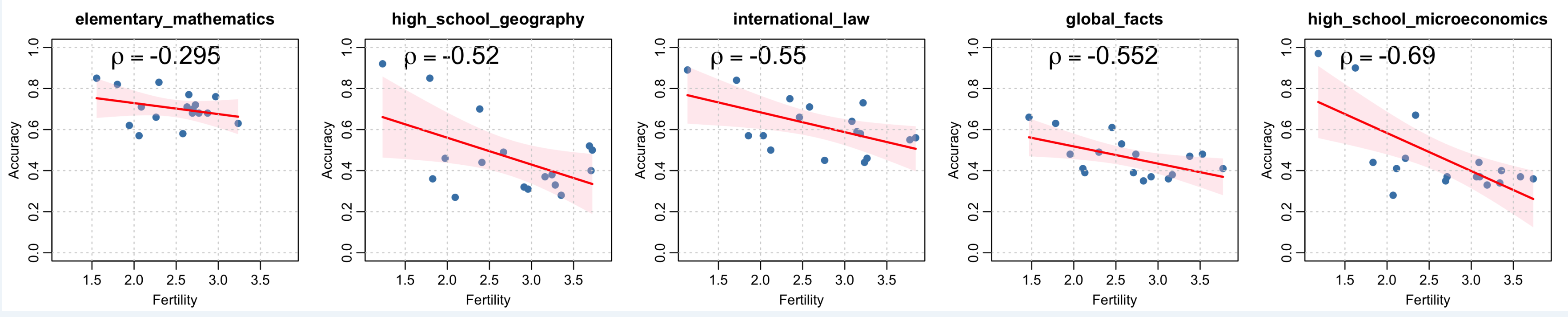


Figure 2. Fertility and accuracy for Llama 3.1 405B across subjects. Strong negative correlations (ρ) demonstrate systematic performance degradation with tokenization inefficiency. Fertility captures tokenization inefficiency that covaries with performance, but does not isolate causal effects independent of pretraining data availability or quality.

Slopes range from -0.08 to -0.18 across all models and subjects. Significant effects after FDR correction include Llama-3.1-405B on Microeconomics (slope $= -0.185$, $p = 0.002$) and Qwen-2.5-32B on Geography (slope $= -0.155$, $p = 0.006$).

References

Adelani et al. (2025). IrokoBench. *NAACL*.

Adebara et al. (2025). Where Are We? Evaluating LLM Performance on African Languages. *ACL*.

Ahia et al. (2023). Do All Languages Cost the Same? *EMNLP*.

Ahia et al. (2024). MAGNET: Multilingual Fairness via Adaptive Gradient-Based Tokenization. *NeurIPS*.

Alhanai et al. (2024). Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages. *arXiv:2412.12417*.

Ali et al. (2024). Tokenizer Choice For LLM Training: Negligible or Crucial? *NAACL Findings*.

Beyene et al. (2025). mSTEB: Massively Multilingual Evaluation of LLMs. *arXiv:2506.08400*.

Joshi et al. (2020). The State and Fate of Linguistic Diversity in the NLP World. *ACL*.

Keles et al. (2022). On The Computational Complexity of Self-Attention. *arXiv:2209.04881*.

NLLB Team (2022). No Language Left Behind. *arXiv:2207.04672*.

Owodunni et al. (2025). FlexiTokens: Flexible Tokenization for Evolving Language Models. *arXiv:2507.12720*.

Petrov et al. (2023). Language Model Tokenizers Introduce Unfairness Between Languages. *arXiv:2305.15425*.

Rust et al. (2021). How Good is Your Tokenizer? *ACL*.

Singh et al. (2025). Global MMLU: Cultural and Linguistic Biases in Multilingual Evaluation. *arXiv:2412.03304*.

Sreedhar et al. (2023). Local Byte Fusion for Neural Machine Translation. *arXiv:2205.11490*.

Vaswani et al. (2017). Attention Is All You Need. *NeurIPS*.