Tokens of Meaning: Predicting Reasoning Efficiency with Information Density

Anonymous ACL submission

Abstract

Reasoning enables large language models to solve complex problems, but practitioners lack guidance for optimizing efficiency before running expensive inference experiments. We introduce an approach that measures information density (tokens per unit of content) and correlates this with reasoning efficiency (accuracy per token). Using 100 parallel passages from FLORES-200 in 50 diverse languages, we measure tokens per passage. We then evaluate reasoning efficiency on 2,696 problems per language across two models and three datasets. Our results reveal a strong positive correlation ($r = 0.933$, $p < 0.0001$) between information density and reasoning efficiency. These findings suggest that higher-density language–tokenizer pairs yield more efficient answers, allowing practitioners to optimize reasoning applications before running expensive evaluation experiments.

1 Introduction

Reasoning has become essential for solving complex problems with large language models. However, there is no free lunch: reasoning can consume significant tokens, increasing latency and costs.

Practitioners face a fundamental question: given a reasoning task, how can they predict which configuration will yield the best accuracy-to-cost trade-off without running exhaustive experiments? Current practice relies on trial and error, benchmarking multiple configurations.

We propose information density as a predictive metric for reasoning efficiency. Our hypothesis is that languages that encode the same semantic content in fewer tokens will produce more efficient reasoning: achieving comparable accuracy with fewer output tokens.

Our contributions are:

1. A methodology for measuring information density across languages using parallel corpora and token counts
2. Empirical evidence from 50 languages and 2 models showing a strong correlation ($r = 0.933$) between information density and reasoning efficiency
3. Practical recommendations for optimizing reasoning-intensive applications based on language selection

2 Background

2.1 Information Theory and Language

Shannon (Shannon, 1948) established the mathematical framework for quantifying information content, showing that the same semantic content can be encoded with varying efficiency while the underlying information remains constant (Cover and Thomas, 2006). This principle has been applied to NLP: Brown et al. (1992) demonstrated that statistical language models can estimate the entropy of English text, and more recently, the perplexity of language models has been used as a proxy for information content (Jurafsky and Martin, 2024). Coupé et al. (2019) measured information rates across 17 human languages, finding that while languages vary dramatically in syllabic rate, the information rate remains remarkably constant at approximately 39 bits/second when spoken, suggesting a universal communication bandwidth.

2.2 Efficient Reasoning in LLMs

Recent work has explored methods for reducing the token cost of reasoning while maintaining accuracy.

Lee et al. (2025) studied 31 compression prompts across 6 LLMs, discovering universal trade-off curves and "token complexity" thresholds below which accuracy degrades. Ahuja et al. (2025) evaluated cross-lingual reasoning across 7

languages, showing that non-English languages can reduce tokens while preserving accuracy. However, they have a target language consistency (TLC — percentage of answers in the desired language) of 0.7–0.8 meaning that a large percentage of questions for each language could have been answered in English. In addition, Petrov et al. (2023) demonstrated that tokenizers introduce significant unfairness across languages, with non-English text requiring up to 15 times more tokens to encode the same content, suggesting that apparent token savings may not generalize across tokenization schemes. Yi et al. (2025) proposed an RL-based method achieving 50–80% length reduction via learned optimal reasoning lengths. Li et al. (2026) introduced "step entropy" to identify redundant reasoning steps, showing 80% of low-entropy steps can be pruned without accuracy loss.

Sui et al. (2025) provided a comprehensive survey categorizing efficient reasoning into model-based, output-based, and prompt-based methods. Gwak et al. (2025) found that successful LLM reasoning exhibits non-uniform information density, unlike human communication patterns. Li et al. (2025) developed THINK-Bench to evaluate reasoning efficiency, revealing that most large reasoning models exhibit overthinking on easy questions.

2.3 Gap in Prior Work

No prior work systematically measures information density across a large set of languages and connects these measurements to reasoning optimization. We provide a measurement-driven framework that can predict reasoning efficiency without expensive benchmark evaluations.

3 Methodology

3.1 Overview

Our methodology consists of two phases:

- **Phase 1 (information density):** Measure tokens per passage across languages using parallel text
- **Phase 2 (reasoning efficiency):** Measure accuracy and token usage on reasoning benchmarks

We then correlate information density with reasoning efficiency.

3.2 Models

We evaluate two models accessed via AWS Bedrock:

- **Claude Sonnet 4.5:** Anthropic's flagship reasoning model
- **Qwen 3 235B:** Alibaba's large multilingual model

Table 3 in the appendix has more details.

3.3 Languages

We selected 50 languages via stratified sampling across three dimensions: writing system (Latin, Arabic, Cyrillic, Indic, CJK, and other scripts), resource level (Joshi et al., 2020, classes 0–5), and language family (12+ families including Romance, Germanic, Slavic, Sino-Tibetan, and Dravidian). Table 5 in the appendix lists all languages with their classifications.

3.4 Phase 1: Information Density Measurement

We randomly sampled 100 passages from the FLORES-200 development split (NLLB Team, 2022), a parallel corpus of 997 passages professionally translated into 200+ languages (released under CC-BY-SA 4.0). For Claude Sonnet 4.5, we prompted the model to repeat each text exactly and recorded the output token count (99.9% of responses matched exactly); for Qwen 3 235B, we used the open-source tokenizer directly. We computed mean tokens per passage for each model and language, defining information density as the inverse of tokens per passage averaged across the 100 samples and finally multiplied by 1,000 for scale. Since FLORES-200 passages are expertly translated to preserve meaning, we assume translations carry equivalent semantic information, so fewer tokens indicates higher information density.

$$\text{Information Density} = \frac{1{,}000}{\text{tokens per passage}} \quad (1)$$

Why not bit-based metrics? An alternative approach would measure bits per token using model perplexity or entropy. We rejected this because bits per token depends heavily on model certainty, biasing the results towards high-resource languages with more training data. Model certainty does not equal information content. Using bits would conflate model quality with language properties.

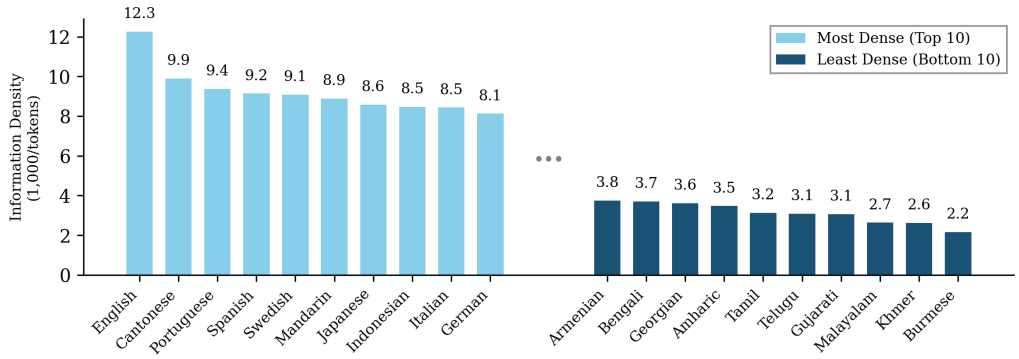

Figure 1: Tokens per passage by language (10 most and 10 least efficient, all models). Dense languages (English, Cantonese) require far fewer tokens than least dense (Burmese, Khmer).

Our token-based approach avoids these confounds by measuring encoding efficiency directly.

Why not semantic similarity metrics? Another approach would use embedding similarity to verify semantic preservation. We rejected this because it requires an arbitrary choice of similarity threshold and different embedding models produce different results. Our approach assumes expert translations preserve meaning, avoiding these issues.

3.5 Phase 2: Reasoning Efficiency Measurement

3.5.1 Datasets

Table 1: Reasoning benchmark datasets used in Phase 2.

Dataset	Examples	Description
MMLU Pro	400	Multi-task understanding
GPQA	448	Graduate-level science
MATH500	500	Competition math
Total	**1,348**	**Per model per language**

We use three challenging benchmarks: MMLU Pro (Wang et al., 2024), GPQA (Rein et al., 2024), and MATH500 (Hendrycks et al., 2021). See table

We translated problems to each target language with a translation model, then back-translated to English, and finally verified the faithfulness with Claude Opus 4.5. If the verification failed, we used a different translation model (Amazon Translate → Claude Opus 4.5 → Qwen3 235b). The success rate was 99.7%.

For each tuple of model, language, dataset, and problem, we:

1. Presented the translated problem with step-by-step reasoning instructions (no specific reasoning budget)
2. Collected the model's response
3. Extracted the final answer and evaluated correctness
4. Counted response tokens
5. Detected response language

The target language consistency (TLC — percentage of answers in the desired language) was 99.8%.

3.5.2 Efficiency Metric

$$\text{Efficiency} = \frac{\text{Accuracy}}{\text{Total Tokens}} \times 1{,}000 \qquad (2)$$

This measures correctness per token spent. Higher values indicate more efficient reasoning.

Addressing tokens in both variables. A potential concern is that tokens appear in both the independent variable and dependent variable. However, the tokens come from different distributions:

- **Independent variable tokens:** From FLORES-200, which consists of human-written, factual, short passages of 1–3 sentences. Measures how efficiently a language–tokenizer encodes information.
- **Dependent variable tokens:** From reasoning tasks, which are model-generated, multi-paragraph reasoning outputs. Measures how concisely a model reasons in language X.

We are using an easy-to-calculate metric (information density) to predict a hard-to-calculate metric (reasoning efficiency).

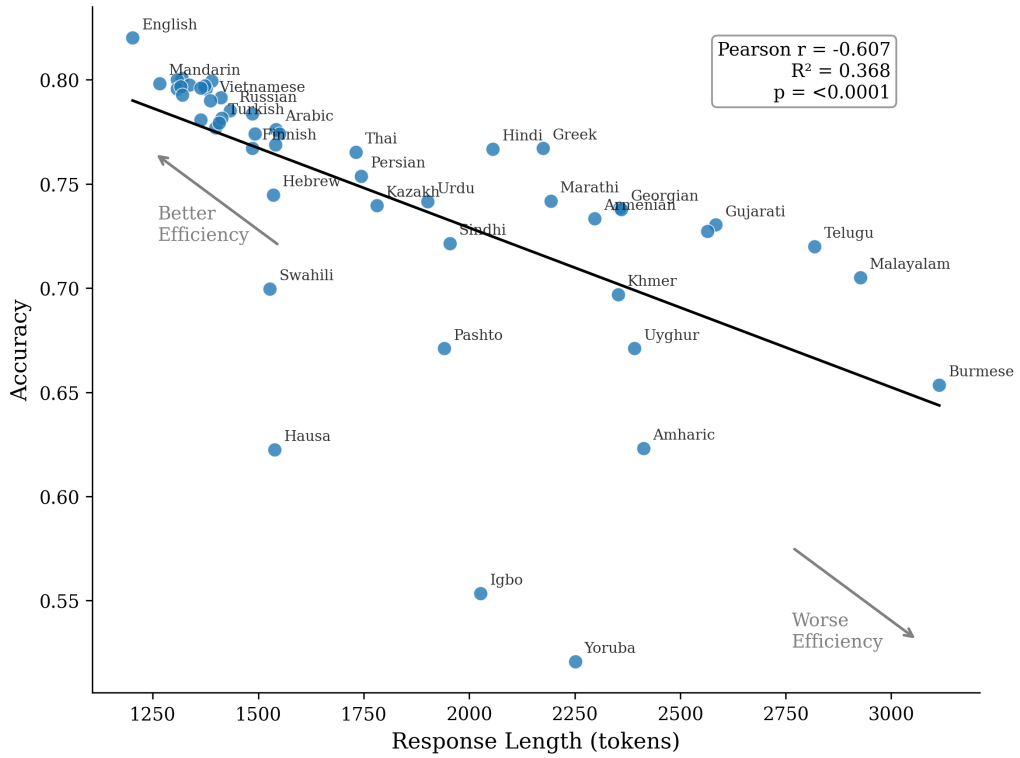

Figure 2: Response length vs. accuracy in Phase 2, with shading for reasoning efficiency. A negative correlation shows that languages producing shorter responses tend to achieve higher accuracy.

4 Results and Analysis

4.1 Sample Sizes

$$\begin{aligned} &50 \text{ lang.} \times 2 \text{ models} \times 1{,}348 \text{ prob.} \\ &= 134{,}800 \text{ evaluations} \end{aligned} \tag{3}$$

4.2 Information Density Results

Figure 1 shows that languages varied substantially in tokens per passage, with an approximately 6× range from most to least dense. This range reflects differences in tokenizer training, script complexity, and language structure. High-resource languages with Latin or Asian scripts tend to be tokenized more efficiently, while languages with less-common scripts or limited representation in tokenizer training data require more tokens.

4.3 Reasoning Efficiency Results

Figure 2 shows the reasoning efficiency results broken down by the numerator (accuracy) and the denominator (response length). Accuracy and response length have very little relationship for low reasoning-efficiency languages, but correlate more tightly for high reasoning-efficiency languages. The correlation is moderate at $r = -0.607$ (and significant $p < 0.0001$). The high-reasoning-efficiency languages are both more accurate and use fewer tokens, showing reasoning efficiency is not simply controlled by tokenizer efficiency across uniform model performance; something else is happening due to model quality in a specific language or language information density.

4.4 Correlation: Information Density and Reasoning Efficiency

Figure 3 shows the central finding: There is a strong positive relationship between information density and reasoning efficiency. The correlation is remarkably strong for a cross-linguistic study. The Pearson coefficient is 0.922 and 85% of the variance in reasoning efficiency can be explained by information density.

4.5 Analysis

Languages that encode information more compactly also produce more efficient reasoning outputs. This relationship holds across diverse language families, writing systems, and resource levels.

Several factors contribute to this pattern. First, as we expect, tokenization efficiency directly affects reasoning length: if language A requires more tokens to express the same reasoning as language

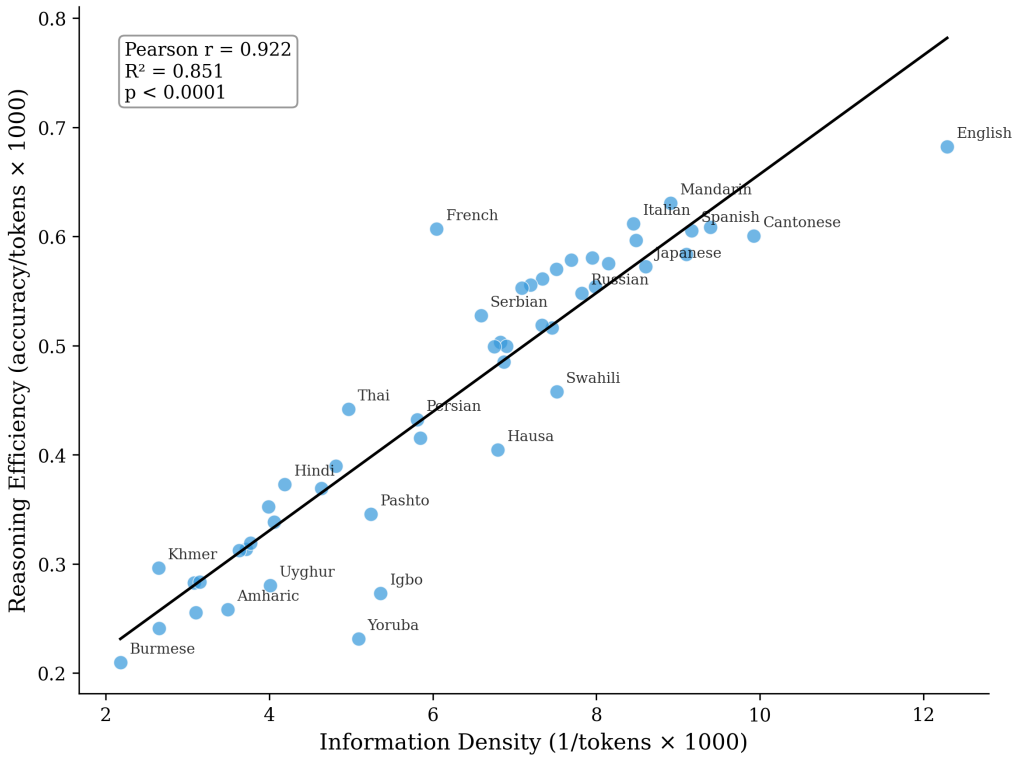

Figure 3: Information density vs. reasoning efficiency with linear regression fit. Languages that encode information more compactly produce more efficient reasoning outputs.

B, the total output grows proportionally. Tokenizers should be more efficient on the more common languages in their training sets. Second, models may have learned more efficient reasoning patterns in languages where they process more text during training.

Notably, when we control for resource level (using Joshi et al., 2020, classification), we see the effect of information density decrease by 36.1% but remain highly significant ($p < .0001$). See Table 2. While high-resource languages do tend to have both better tokenization and better model performance, the relationship between information density and reasoning efficiency cannot be fully explained by language resource level.

The practical implication is clear: information density measured on simple parallel text can predict relative reasoning efficiency on complex benchmarks. This enables practitioners to estimate efficiency tradeoffs without running expensive evaluations.

5 Discussion

5.1 Claims Supported by Evidence

Our results support the following claims:

Table 2: OLS regression: reasoning efficiency.

	Bivariate	With Controls
Info. Density	0.0545***	0.0348***
	(0.0033)	(0.0049)
Intercept	0.1127***	0.1731***
	(0.0218)	(0.0283)
Joshi Controls	—	
Class 2	—	−0.0063 (0.0213)
Class 3	—	0.0489** (0.0163)
Class 4	—	0.0643** (0.0184)
Class 5	—	0.0777** (0.0222)
R^2	0.8509	0.9490
Adj. R^2	0.8478	0.9324
F-statistic	273.88***	57.33***

SE in parentheses. *p<.05, **p<.01, ***p<.001
N=50. Efficiency scaled ×1000.

1. **Information density correlates strongly with reasoning efficiency** ($r = 0.933$)**.** Languages with more compact token representations produce more efficient reasoning outputs.

2. **Practical recommendations:**

   - For applications serving users in low-density languages: consider performing behind-the-scenes reasoning in a high-

density language, then translating results.

- For language-agnostic applications: prefer high-density reasoning languages to minimize costs.
- When comparing reasoning approaches: favor approaches with higher information density as a proxy for efficiency without running full evaluations.

5.2 Claims Not Supported

We do not claim:

1. Certain languages have "naturally" higher information density. Our measurements reflect specific models' tokenizers and training. Different models may lead to different language rankings.

2. Information density *causes* better reasoning efficiency. Our correlation does not establish causation. A third factor, such as model quality or training data quality for each language, could drive both metrics.

6 Conclusion

We introduced information density as a predictive metric for reasoning efficiency in large language models. By measuring tokens per passage on parallel text and correlating with reasoning performance on benchmark tasks, we found a strong positive correlation ($r = 0.933$) across 50 languages and 2 models.

These findings provide practitioners with evidence-based guidance:

- Information density can predict reasoning efficiency without running expensive benchmarks

- High-density languages offer efficiency advantages

- When optimizing reasoning costs, consider language selection alongside model and prompt engineering

Future work should extend these findings to more models, investigate causal mechanisms, and explore whether prompting strategies that increase information density yield similar efficiency gains.

Limitations

Our study has several limitations that should be considered when interpreting the results:

- **Two models only:** We evaluated two LLMs (Claude Sonnet 4.5 and Qwen 3 235B). Results may not generalize to all model architectures, sizes, or tokenizer designs.

- **Translation quality:** Although we employed a multi-stage translation and verification pipeline (with 99.7% success rate), some problems may have lost nuance or domain-specific meaning in translation, particularly for low-resource languages.

- **Tokenizer-specific measurements:** Information density is inherently tied to a specific tokenizer. Different tokenizers would yield different language rankings, so our results should not be interpreted as intrinsic properties of languages.

- **Correlation vs. causation:** While the correlation is strong ($r = 0.933$), we do not establish a causal mechanism. A confounding variable, such as training data volume, could drive both information density and reasoning quality.

- **English-centric benchmarks:** Our reasoning benchmarks (MMLU Pro, GPQA, MATH500) originate in English. Languages with different conceptual frameworks or knowledge traditions may be systematically disadvantaged by translated versions of these benchmarks.

- **Static evaluation:** We evaluated at a single point in time. As models and tokenizers evolve, the specific rankings and correlation strength may change, though the general relationship is likely to persist.

Ethical Considerations

Our findings carry equity implications. Current tokenizers introduce significant unfairness across languages, with non-English text requiring up to 15× more tokens to encode equivalent content. Speakers of low-density languages—particularly those using non-Latin, non-CJK scripts such as Burmese, Khmer, Malayalam, and Telugu—may be further disadvantaged if practitioners apply our recommendations uncritically, potentially widening existing gaps for already underserved communities. The recommendation to reason in high-density

languages and translate results also introduces error risk in sensitive domains such as healthcare and legal applications, and because our findings are tokenizer-specific, applying rankings from one model to another could yield incorrect decisions.

Our methodology inherently reflects these tokenizer biases, and our English-origin benchmarks may disadvantage languages with different conceptual frameworks. Practitioners should interpret our rankings as reflecting the current state of tokenizer development rather than fixed properties of languages themselves.

Use of AI Assistants

We used AI writing assistants (including large language models) to help draft, edit, and format this paper, always under direct human supervision. The authors take full responsibility for all content in this paper.

References

Sanchit Ahuja, Preetham Vaddamanu, and Barun Patra. 2025. EfficientXLang: Towards improving token efficiency through cross-lingual reasoning. *arXiv preprint arXiv:2507.00246*.

Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2nd edition. Wiley-Interscience.

Minjae Gwak, Giwon Son, and Jungwoo Kim. 2025. Revisiting the uniform information density hypothesis in LLM reasoning traces. *arXiv preprint arXiv:2510.13850*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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)*, pages 6282–6293.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing*, 3rd edition draft edition. Stanford University.

Andrew Lee, Edoardo Che, and Tianyu Peng. 2025. How well do LLMs compress their own chain-of-thought? A token complexity approach. *arXiv preprint arXiv:2503.01141*.

Zhen Li, Yang Chang, and Yifan Wu. 2025. THINK-Bench: Evaluating thinking efficiency and chain-of-thought quality of large reasoning models. *arXiv preprint arXiv:2505.22113*.

Zhuohan Li, Jieyu Zhong, Ziming Zheng, and 1 others. 2026. Making slow thinking faster: Compressing LLM chain-of-thought via step entropy. In *International Conference on Learning Representations (ICLR)*.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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)*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level Google-proof Q&A benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Yuan Sui, Yu-Neng Chuang, Guanchu Wang, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. Technical report, Rice University.

Yubo Wang, Xueguang Ma, Ge Zhang, and 1 others. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Jie Yi, Jian Wang, and Shuo Li. 2025. ShorterBetter: Guiding reasoning models to find optimal inference length for efficient reasoning. *arXiv preprint arXiv:2504.21370*.

A Model Details

Table 3: Model specifications and computational cost. All experiments used temperature = 0 and max output tokens = 4,096 (well above typical response lengths). Models accessed via AWS Bedrock.

Model	Parameters	Model ID	Cost
Claude Sonnet 4.5	∼175B	`claude-sonnet-4-5-*`	$948.80
Qwen 3 235B	235B	`qwen3-vl-235b-a22b`	$115.70
Total			$1,064.50

Cost breakdown: Phase 1 (tokenization): $8.13; Phase 2 (reasoning): $1,056.37. Total API calls: ∼270k.

B Artifact Licenses and Intended Use

Table 4: Licenses and intended use for all artifacts.

Artifact	License	Intended Use
FLORES-200	CC-BY-SA 4.0	Multilingual evaluation
MMLU Pro	MIT	Multi-task language understanding evaluation
GPQA	MIT	Graduate-level reasoning evaluation
MATH500	MIT	Mathematical reasoning evaluation

All existing artifacts were used consistent with their intended purposes: FLORES-200 was designed for multilingual evaluation and we use it for cross-lingual token measurement; MMLU Pro, GPQA, and MATH500 are reasoning benchmarks and we use them for reasoning evaluation. No license terms were violated.

Our translated benchmark derivatives are used solely for research purposes, consistent with the original access conditions. We do not redistribute the translated datasets.

C Language Classifications

Table 5: Language classification: 50 languages stratified by writing system, family, and resource level.

Language	ISO	Script	Family	Joshi
Amharic	amh	Other	Afroasiatic	2
Arabic	arb	Arabic	Afroasiatic	5
Armenian	hye	Other	Armenian	1
Bengali	ben	Indic	Indo-Aryan	3
Bulgarian	bul	Cyrillic	Slavic	3
Burmese	mya	Other	Sino-Tibetan	1
Cantonese	yue	CJK	Sino-Tibetan	1
Chinese	zho	CJK	Sino-Tibetan	5
Czech	ces	Latin	Slavic	4
English	eng	Latin	Germanic	5
Finnish	fin	Latin	Uralic	4
French	fra	Latin	Romance	5
Georgian	kat	Other	Kartvelian	3
German	deu	Latin	Germanic	5
Greek	ell	Other	Hellenic	3
Gujarati	guj	Indic	Indo-Aryan	1
Hausa	hau	Latin	Afroasiatic	2
Hebrew	heb	Other	Afroasiatic	3
Hindi	hin	Indic	Indo-Aryan	4
Hungarian	hun	Latin	Uralic	4
Igbo	ibo	Latin	Niger-Congo	1
Indonesian	ind	Latin	Austronesian	3
Italian	ita	Latin	Romance	4
Japanese	jpn	CJK	Japonic	5
Kazakh	kaz	Cyrillic	Turkic	3
Khmer	khm	Other	Austroasiatic	1
Korean	kor	CJK	Koreanic	4
Malayalam	mal	Indic	Dravidian	1
Marathi	mar	Indic	Indo-Aryan	2
Pashto	pbt	Arabic	Iranian	1
Persian	pes	Arabic	Iranian	4
Polish	pol	Latin	Slavic	4
Portuguese	por	Latin	Romance	4
Romanian	ron	Latin	Romance	3
Russian	rus	Cyrillic	Slavic	4
Serbian	srp	Cyrillic	Slavic	4
Sindhi	snd	Arabic	Indo-Aryan	1
Spanish	spa	Latin	Romance	5
Swahili	swh	Latin	Niger-Congo	2
Swedish	swe	Latin	Germanic	4
Tagalog	tgl	Latin	Austronesian	3
Tamil	tam	Indic	Dravidian	3
Telugu	tel	Indic	Dravidian	1
Thai	tha	Other	Tai-Kadai	3
Turkish	tur	Latin	Turkic	4
Ukrainian	ukr	Cyrillic	Slavic	3
Urdu	urd	Arabic	Indo-Aryan	3
Uyghur	uig	Arabic	Turkic	1
Vietnamese	vie	Latin	Austroasiatic	4
Yoruba	yor	Latin	Niger-Congo	2

D Phase 1: Information Density by Language

Table 6: Phase 1: Information density measurements (100 FLORES-200 passages per language, both models averaged). Values show mean ± 95% CI.

Language	Tokens per Passage	Information Density
Amharic	286.1(23.7)	3.496(0.290)
Arabic	146.5(8.1)	6.827(0.377)
Armenian	265.1(18.3)	3.772(0.260)
Bengali	268.9(13.1)	3.719(0.181)
Bulgarian	144.9(8.3)	6.900(0.393)
Burmese	458.6(36.6)	2.181(0.174)
Cantonese	100.8(4.5)	9.924(0.447)
Chinese	112.2(9.0)	8.909(0.714)
Czech	136.3(7.0)	7.337(0.377)
English	81.4(2.7)	12.289(0.409)
Finnish	134.1(5.5)	7.455(0.306)
French	165.4(25.6)	6.047(0.934)
Georgian	275.4(14.7)	3.631(0.194)
German	122.8(7.0)	8.145(0.466)
Greek	250.6(13.1)	3.991(0.209)
Gujarati	324.6(24.7)	3.080(0.234)

continued on next page

Table 6 – *continued*

Language	Tokens per Passage	Information Density
Hausa	147.2(10.3)	6.796(0.478)
Hebrew	145.6(10.5)	6.869(0.495)
Hindi	238.7(12.6)	4.190(0.222)
Hungarian	136.3(7.3)	7.335(0.394)
Igbo	186.5(14.5)	5.361(0.417)
Indonesian	117.8(6.1)	8.487(0.442)
Italian	118.3(5.6)	8.453(0.400)
Japanese	116.2(4.8)	8.605(0.354)
Kazakh	171.0(7.1)	5.848(0.244)
Khmer	377.9(17.6)	2.646(0.124)
Korean	138.9(8.4)	7.197(0.434)
Malayalam	376.5(19.2)	2.656(0.136)
Marathi	246.4(11.0)	4.059(0.181)
Pashto	190.7(9.7)	5.244(0.268)
Persian	172.2(7.5)	5.808(0.254)
Polish	130.0(7.6)	7.695(0.452)
Portuguese	106.5(4.3)	9.393(0.379)
Romanian	125.8(6.4)	7.948(0.407)
Russian	127.9(5.5)	7.819(0.336)
Serbian	151.7(9.5)	6.590(0.413)
Sindhi	215.6(17.3)	4.639(0.373)
Spanish	109.1(3.7)	9.165(0.314)
Swahili	133.1(6.3)	7.515(0.353)
Swedish	109.9(4.4)	9.101(0.368)
Tagalog	141.1(7.6)	7.086(0.380)
Tamil	317.4(16.9)	3.151(0.168)
Telugu	322.3(17.6)	3.103(0.170)
Thai	201.2(10.9)	4.971(0.268)
Turkish	125.2(5.4)	7.985(0.342)
Ukrainian	148.2(6.9)	6.749(0.314)
Urdu	207.6(10.2)	4.816(0.237)
Uyghur	249.3(12.5)	4.012(0.201)
Vietnamese	133.2(7.5)	7.510(0.423)
Yoruba	196.4(14.9)	5.093(0.386)

E Phase 2: Reasoning Performance by Language

Table 7: Phase 2: Reasoning performance (2,696 problems per language, both models averaged). Values show mean ± 95% CI.

Language	Accuracy	Mean Resp. Tokens	TLC	Reasoning Eff.
Amharic	0.623(0.018)	2413.6(50.9)	0.995(0.003)	0.258(0.012)
Arabic	0.776(0.016)	1542.1(36.0)	0.998(0.002)	0.503(0.018)
Armenian	0.733(0.017)	2297.1(51.9)	0.998(0.002)	0.319(0.012)
Bengali	0.739(0.017)	2355.9(48.5)	1.000	0.313(0.011)
Bulgarian	0.774(0.016)	1549.7(35.9)	0.999(0.001)	0.500(0.018)
Burmese	0.654(0.018)	3113.8(62.0)	0.997(0.002)	0.210(0.008)
Cantonese	0.793(0.015)	1320.1(34.7)	0.998(0.002)	0.600(0.023)
Chinese	0.798(0.015)	1265.9(32.9)	0.999(0.001)	0.631(0.024)
Czech	0.792(0.015)	1410.8(34.7)	0.997(0.002)	0.561(0.021)
English	0.820(0.015)	1202.1(37.2)	1.000	0.682(0.028)
Finnish	0.767(0.016)	1485.7(36.0)	0.998(0.002)	0.516(0.020)
French	0.801(0.015)	1319.4(34.1)	0.999(0.001)	0.607(0.023)
Georgian	0.738(0.017)	2360.6(50.5)	0.999(0.001)	0.313(0.012)
German	0.799(0.015)	1389.9(33.8)	0.999(0.001)	0.575(0.021)
Greek	0.767(0.016)	2175.1(45.6)	0.997(0.002)	0.353(0.012)
Gujarati	0.730(0.017)	2584.5(56.0)	0.999(0.001)	0.283(0.011)
Hausa	0.622(0.018)	1538.9(43.9)	0.996(0.002)	0.404(0.019)
Hebrew	0.745(0.016)	1535.1(38.4)	0.996(0.002)	0.485(0.020)
Hindi	0.767(0.016)	2056.1(45.9)	0.998(0.002)	0.373(0.014)
Hungarian	0.774(0.016)	1492.4(35.8)	0.997(0.002)	0.519(0.019)
Igbo	0.553(0.019)	2027.0(52.8)	0.990(0.004)	0.273(0.014)
Indonesian	0.797(0.015)	1337.0(33.7)	0.997(0.002)	0.596(0.022)
Italian	0.800(0.015)	1307.7(34.7)	0.999(0.001)	0.612(0.023)
Japanese	0.781(0.016)	1363.8(31.2)	0.999(0.001)	0.572(0.021)
Kazakh	0.740(0.017)	1780.8(38.9)	0.997(0.002)	0.415(0.016)
Khmer	0.697(0.017)	2352.7(47.6)	0.997(0.002)	0.296(0.011)
Korean	0.777(0.016)	1398.4(29.0)	0.999(0.001)	0.556(0.019)
Malayalam	0.705(0.017)	2926.9(55.7)	0.999(0.001)	0.241(0.009)
Marathi	0.742(0.017)	2193.0(45.8)	0.998(0.002)	0.338(0.012)
Pashto	0.671(0.018)	1940.8(45.0)	0.998(0.002)	0.346(0.015)
Persian	0.754(0.016)	1744.0(36.4)	0.999(0.001)	0.432(0.015)
Polish	0.796(0.015)	1375.7(34.5)	0.998(0.002)	0.578(0.022)
Portuguese	0.796(0.015)	1307.4(34.9)	0.999(0.001)	0.609(0.023)
Romanian	0.797(0.015)	1373.4(34.6)	0.998(0.002)	0.580(0.022)
Russian	0.785(0.016)	1432.6(35.7)	0.997(0.002)	0.548(0.021)
Serbian	0.784(0.016)	1485.9(37.7)	0.999(0.001)	0.527(0.020)
Sindhi	0.721(0.017)	1953.7(42.8)	0.997(0.002)	0.369(0.015)

continued on next page

Table 7 – *continued*

Language	Accuracy	Mean Resp. Tokens	TLC	Reasoning Eff.
Spanish	0.797(0.015)	1316.0(34.1)	0.998(0.002)	0.605(0.023)
Swahili	0.700(0.017)	1527.5(38.8)	0.996(0.002)	0.458(0.020)
Swedish	0.796(0.015)	1363.9(36.0)	0.998(0.002)	0.584(0.022)
Tagalog	0.782(0.016)	1413.3(33.2)	0.996(0.002)	0.553(0.020)
Tamil	0.727(0.017)	2564.4(52.4)	0.999(0.001)	0.284(0.010)
Telugu	0.720(0.017)	2818.8(59.7)	0.999(0.001)	0.255(0.010)
Thai	0.765(0.016)	1731.5(33.5)	0.999(0.001)	0.442(0.015)
Turkish	0.779(0.016)	1406.7(34.5)	0.999(0.001)	0.554(0.021)
Ukrainian	0.769(0.016)	1540.6(36.8)	0.997(0.002)	0.499(0.019)
Urdu	0.741(0.017)	1901.7(41.4)	0.999(0.001)	0.390(0.015)
Uyghur	0.671(0.018)	2391.9(43.7)	0.997(0.002)	0.281(0.011)
Vietnamese	0.790(0.015)	1385.9(30.3)	0.998(0.002)	0.570(0.020)
Yoruba	0.521(0.019)	2251.4(48.1)	0.997(0.002)	0.231(0.012)