Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3159

TOKEN METROLOGY FOR LARGE LANGUAGE MODELS: A SURVEY FROM THE PERSPECTIVE OF MEASUREMENT SCIENCE

YingYu Lin*, YuJie Liufu
Quality and Metrology Supervision and Testing Institute, Maoming 525000, Guangdong, China.
**Corresponding Author: YingYu Lin*

Abstract: The token has become the de facto universal unit of measurement in the large language model (LLM) ecosystem: model pricing is quoted per token, context windows are expressed in tokens, inference throughput is measured in tokens per second, and training corpora are reported in trillions of tokens. Yet the token is not a well-defined measurement unit in the metrological sense—it varies with the tokenizer, the vocabulary, text normalization, and the vendor's counting convention; it is incommensurable across models; and it lacks calibration, traceability, and an assessable uncertainty. This paper examines token metering as an object of measurement science. We first survey the state of token counting, covering the lineage of tokenization techniques, multilingual and domain disparities, divergent billing conventions, and reproducibility issues, substantiated by a first-hand benchmark across 13 tokenizers and 18 multilingual parallel texts. Drawing on the vocabulary of metrology (the International Vocabulary of Metrology and the Guide to the Expression of Uncertainty in Measurement), we then establish a four-dimensional framework—definiteness, commensurability, traceability, and uncertainty—through which we identify four systemic deficiencies of token counting as a measurement system. We argue that these deficiencies undermine billing fairness, cost comparability, and scientific reproducibility. Finally, we review existing mitigation efforts and outline a standardization agenda spanning technical, metrological, and scholarly-reporting levels. We conclude that token metering currently occupies a "pre-SI" stage of measurement, and that establishing a public reference standard and measurement norms is a prerequisite for the LLM economy and for reproducible LLM science.
Keywords: Large language models; Tokenization; Metrology; Billing; Measurement uncertainty; Calibration and traceability

1 INTRODUCTION

Since 2018, Transformer-based large language models (LLMs) have become a foundational infrastructure of artificial intelligence, and one concept is nearly ubiquitous: the token. Pricing pages quote costs "per million tokens"; context windows are specified in token counts; inference throughput is reported in tokens per second; training teams describe corpus size in "X trillion tokens"; and even the perplexity and floating-point operations (FLOPs) reported in the academic literature are normalized by token count. It is not an overstatement that the token has become the most frequently used and most consequential "unit of measurement" of the LLM era.

However, when examined through the lens of measurement science—metrology—the meaning of "one token" remains disturbingly ill-defined. The same text yields different token counts under the byte-pair-encoding (BPE) tokenizer of the GPT-2 era and the o200k tokenizer of GPT-4o, sometimes differing by a factor of several (our measurements in Section 3.2 show that a 71-character Chinese passage is counted nearly three times larger by the former than by the latter). Different models of the same vendor use vocabularies ranging from 50,000 to 200,000 entries. The same text produces different counts under different Unicode normalization forms. More critically, vendors differ in which tokens enter billing: whether input and output are priced separately, whether cache-hit tokens are discounted, and whether hidden tokens introduced by system prompts and tool calls are charged. Users can hardly audit their bills; researchers can hardly reproduce reported token counts; and the widely used comparison metric "price per million tokens" is, strictly speaking, not meaningful across models—because a token of one model is not the same unit as a token of another.

This situation is not unfamiliar in the history of measurement. Before the metric system, European "feet" and "pounds" were defined differently across jurisdictions; the International System of Units (SI) was established precisely to provide well-defined units, reproducible realization methods, traceable calibration chains, and assessable uncertainties [1]. Token metering today occupies a similar "pre-SI" stage: units are defined unilaterally by vendors, measurement procedures are opaque or mutually inconsistent, no public reference standard exists, and the uncertainty of the reported counts cannot be evaluated.

This paper makes four contributions. First, we systematically survey the state of token counting (Sections 2–3) and report a first-hand, reproducible benchmark across 13 tokenizers and 18 multilingual parallel texts. Second, drawing on the VIM/GUM conceptual apparatus, we establish a four-dimensional framework—definiteness, commensurability, traceability, and uncertainty—and consolidate the scattered practical problems into four systemic deficiencies (Section 4). Third, we analyze the consequences of these deficiencies for billing fairness and scientific reproducibility

© By the Author(s) 2026, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

(Sections 4.5 and 4.6). Fourth, we review existing mitigation efforts and articulate a standardization agenda across the technical, metrological, and scholarly-reporting levels (Section 5), followed by conclusions and limitations (Sections 6–7).

2 BACKGROUND: TOKENIZATION AND METROLOGY

2.1 Tokenization Techniques

A token is the elementary unit produced by a tokenizer when segmenting text. Early neural language models operated on characters or whole words, but character-level modeling suffered from excessive sequence length and word-level modeling from data sparsity and out-of-vocabulary words. Since 2016, subword tokenization has become dominant; the lineage is surveyed by Mielke et al. [2]. The most influential scheme is byte-pair encoding (BPE), which starts from characters and iteratively merges the most frequent adjacent fragments in a corpus, yielding a merge tree, a fixed-size vocabulary, and a greedy segmentation rule [3].

The principal schemes and their representative models are as follows:

BPE: introduced by Sennrich et al. [3]; GPT-2 uses a 50,257-entry BPE vocabulary [4]; OpenAI's later cl100k_base (about 100,000 entries) and o200k_base (about 200,000 entries) are engineering realizations of BPE serving GPT-3.5/GPT-4 and GPT-4o onward, respectively [5].

WordPiece: a Google variant that replaces the "most frequent merge" criterion with "maximum likelihood merge," used by BERT [6].

SentencePiece / Unigram: a language-independent framework by Kudo and Richardson that models raw text (including whitespace), adopted by LLaMA 1/2 and Gemini, among others [7-8].

Two features are metrologically salient. First, a tokenizer is a product of its training corpus: the vocabulary and merge rules are determined by corpus statistics, so the token count of a given text depends heavily on the corpus on which the tokenizer was trained. Tokenizers trained predominantly on English segment Chinese, Japanese, and similar languages substantially less efficiently. Second, the vocabulary contains special tokens—end-of-text markers, BOS/EOS markers, padding tokens—whose treatment in counting, billing, and context-window accounting varies across vendors, contributing to divergent counting conventions.

2.2 Fundamentals of Metrology

Metrology is "the science of measurement and its application" [9]. The International Vocabulary of Metrology (VIM) and the Guide to the Expression of Uncertainty in Measurement (GUM) provide the core concepts of modern measurement, which can be organized into four pillars, each corresponding to one deficiency identified in Section 4 [9-10]:

Definiteness of the measurand (corresponding to Deficiency I): the measurand must be specified before measurement. An ill-defined measurand renders the measurement result meaningless.

Unit and reproducibility (corresponding to Deficiency II): measurement results are expressed in units, and a unit must have an operational definition and a reproducible realization. The SI units are realized through defining physical constants, so that any competent laboratory reproduces the same result [1].

Calibration and traceability (corresponding to Deficiency III): measuring instruments are calibrated, and the calibration chain is traced back, step by step, to reference standards and ultimately to SI units, with the uncertainty recorded at every step to form a traceability chain. Standard reference materials/data are the physical carriers of calibration.

Uncertainty (corresponding to Deficiency IV): every measurement result should be accompanied by an uncertainty—a quantified expression of the reliability of the result [10].

Measured against these four pillars, token counting exhibits the following state: the measurand depends on the measuring instrument; the unit is not reproducible across vendors; no calibration chain or reference standard exists; and the uncertainty cannot be evaluated. Section 4 elaborates each point. We emphasize that our appeal to metrology is analogical: we import its requirement of measurement-system completeness (definable, reproducible, calibratable, auditable) rather than claiming that the token should be "defined" in the ontological sense of an SI unit.

2.3 Roles of Token Counting in the LLM Ecosystem

Token counting serves multiple measurement functions in the LLM ecosystem, each of which imposes consistency requirements:

Billing metrology: API vendors charge by token count, with separate prices for input, output, and cache hits [11-13]. This is economic metrology, directly affecting money flows.

Capacity metrology: context-window length and rate limits (tokens per minute) are expressed in tokens.

Performance metrology: inference throughput (tokens per second) and per-token latency are core service metrics.

Scale metrology: training corpora are counted in tokens, and training-cost accounting (FLOPs/token) uses the token as the denominator [14-15].

Scholarly metrology: perplexity is normalized by token, and benchmarks report input lengths and costs in token counts.

These functions share the same "token unit" yet demand the same consistency of its definition; when the unit itself diverges, all downstream measurements are distorted in lockstep.

2.4 Terminology

For terminological consistency, we define the key terms of this paper as follows:
Token metering (token metrology): the study of token counting as a measurement process, examining the definiteness of the measurand, the unit, reproducibility, calibration, traceability, and uncertainty.
Tokenizer / vocabulary: the segmentation algorithm and its frozen merge rules and symbol set.
Billing convention: the vendor's rules determining which tokens enter billing and at what price, such as input/output splitting, cache-hit discounts, and the treatment of hidden tokens.
Token tax (tokenizer tax): the phenomenon whereby differences in segmentation efficiency cause certain languages or texts to be systematically over-counted in token number and cost (also termed tokenizer aging or script tax in the literature; we use "token tax" throughout).
Measurand, calibration, traceability, uncertainty: used in the senses of VIM/GUM [9-10].

3 THE STATE OF TOKEN COUNTING

3.1 Divergent Counting Methods: One Token per Vendor

The first and foremost fact about token metering is that no unified token definition exists; what exists are mutually independent tokenizers and vocabularies. Table 1 summarizes the tokenizers and vocabulary sizes of major model families. Even when vocabulary sizes are comparable, differing merge rules and training corpora still produce different segmentations of the same text.
Vocabulary sizes in Table 1 are taken from the respective technical reports and public documentation (e.g., the DeepSeek-V3 technical report [16]). A widely used but strictly invalid rule of thumb follows: "one token equals roughly 0.75 English words, or about four English characters" [5]. The rule holds approximately only for English prose that resembles the tokenizer's training distribution; the error grows rapidly once the text deviates—for code, mathematical formulas, non-English languages, or emoji. For instance, Chinese text typically consumes 0.6–2 tokens per character, depending on the tokenizer (see the measurements in Section 3.2). In short, no stable conversion ratio exists between tokens and text length; any single coefficient is merely a fit to a particular corpus distribution.

Table 1 Tokenizers and Vocabulary Sizes of Major Models

Model/family	Tokenizer	Algorithm	Vocab size (approx.)	Open
GPT-2	gpt2	BPE	50,257	Yes
GPT-3.5 / GPT-4	cl100k_base	BPE	100,000	Yes
GPT-4o onward	o200k_base	BPE	200,019	Yes
BERT	WordPiece	WordPiece	30,522	Yes
LLaMA 1 / 2	SentencePiece	BPE	32,000	Yes
LLaMA 3	LLaMA-3 tokenizer	BPE	128,256	Yes (gated)
Qwen2 / 2.5	Qwen tokenizer	BPE	151,643	Yes
DeepSeek-V3	DeepSeek tokenizer	BPE	128,256	Yes

3.2 Multilingual and Domain Disparity: Same Content, Different "Price"

The dependence of tokenizers on corpus distribution is most pronounced across languages and is the most thoroughly studied aspect of token metering. A line of work has quantified the "multilingual token tax": Ahia et al. introduced the notion of *tokenizer aging* and systematically measured the segmentation efficiency of mainstream tokenizers across languages, finding that languages other than English are generally segmented less efficiently [17]; Deng et al. showed that segmentation inefficiency translates not only into cost and latency but also into a double burden in energy and climate cost, placing non-English speakers at an economic and utility disadvantage [18]; tokenizer choice itself has been shown to materially affect model training and downstream performance [19]; and recent systematic studies characterize this bias as a "token tax" and as "infrastructure bias" [20-21].
To provide first-hand evidence for these qualitative findings, we measured token counts across 13 tokenizers using tiktoken and tokenizers (Table 2 is a representative excerpt; the complete multilingual benchmark is in Appendix A).

Table 2 Token Counts of the Same Test Text under Different Tokenizers. Parenthetical Values Denote Vocabulary Sizes

Segment	GPT-2	cl100k	o200k	LLaMA-3	Qwen2.5	DeepSeek-V3
	(50,257)	(100,277)	(200,019)	(128,256)	(151,665)	(128,815)
English prose (29 words)	33	33	33	34	33	33
Chinese prose (71 chars)	141	71	50	56	46	40
Code (≈100 bytes)	39	23	23	24	23	24
Math (≈56 bytes)	34	35	35	36	35	33
Emoji (4)	13	13	8	14	9	10
Chinese token/char	1.99	1.00	0.70	0.79	0.65	0.56

Table 2 reveals three facts. First, English prose is nearly invariant across six tokenizers (33–34 tokens): texts near the training distribution are counted consistently. Second, the Chinese segment differs by a factor of 3.5: the same 71-character text is counted as 141 tokens in the GPT-2 era but only 40 tokens by DeepSeek-V3; even among contemporary models, o200k (50) and DeepSeek-V3 (40) differ by 25%. Third, the Chinese token/char ratio falls from 1.99 to 0.56, showing that segmentation efficiency is a historical, content-dependent variable of "tokenizer × training corpus," not an intrinsic property of the text. These data directly support the argument of Section 4: token count is not an intrinsic property of a text but a joint function of "text × tokenizer × training corpus."

Extending this illustrative measurement to a larger multilingual benchmark (13 tokenizers × 18 segments, comprising the parallel texts of Article 1 of the Universal Declaration of Human Rights in 12 languages plus technical, code, mathematical, and symbolic samples; see Appendix A for methods and data) yields a systematic conclusion: token inflation grows with the distance between the writing system and the Latin alphabet. Taking English as the baseline, Western European languages (German, Spanish, French) consume 1.32–1.48× the tokens of English; Russian 1.89×; Japanese 2.17×, Arabic 2.29×, Korean 2.69×, Vietnamese 2.72×; Thai 3.78×; and Hindi as much as 5.04×. More importantly, the efficiency of the same language varies by a factor of 7–12 across tokenizers (Hindi 1.61–12.15×; Thai 1.58–10.52×), demonstrating that "how many tokens a language consumes" is itself a quantity that fluctuates sharply with the tokenizer.

3.3 Divergent Billing Conventions: What Is Counted, and at What Price

Even if one accepts that tokens are defined by individual tokenizers, billing conventions diverge further, rendering "price per million tokens" nearly meaningless for cross-vendor comparison. Table 3 summarizes the differences.

Table 3 Token Billing Conventions of Major Vendors (Subject to Official Updates)

Vendor	Unit	Input/output	Cache discount	Hidden tokens	Counter
OpenAI	token	output > input	≈90% cached-input	included	tiktoken
Anthropic	token	output > input	prompt-caching	included	official docs
DeepSeek	token	output > input	cache-hit, peak/off-peak	included	official docs
Baidu Qianfan (ERNIE)	chars	input+output chars	—	—	char-based

Several systemic phenomena follow. First, input/output splitting makes the cost of a request with the same "token count" vary by several times depending on the input/output ratio. Second, cache-hit discounts make the bill for "the same request and the same input" depend on cache state and call time, so that billing is no longer a function of the text alone. Third, the rules for hidden tokens (system prompts, tool calls, special tokens) vary across vendors. Fourth, counters and servers may disagree: the local open-source tokenizer and the server-side count can drift. Notably, a few platforms (such as the early ERNIE series of Baidu Qianfan) billed by character count rather than token count [22]—itself an evasion of the incommensurability defect of the token unit, yet one that further fragments the billing unit.

3.4 Reproducibility Issues

Even with the tokenizer and vocabulary fixed, token counting is not automatically reproducible: Unicode normalization forms (NFC vs. NFD produce different byte sequences for the same text), newline and whitespace handling, emoji zero-width-joiner sequences, and library-version differences all alter the count. These details are almost never reported in papers, benchmarks, or commercial bills. Section 4.4 analyzes the metrological consequence of this phenomenon; here we only record the fact that the same text routinely yields different token counts under different tools, versions, and normalization treatments.

4 A METROLOGICAL ANALYSIS

The phenomena of Section 3, if treated as mere engineering matters, would be dismissed as "different vendors, different implementations." This section instead imports the metrological framework of Section 2 and elevates the scattered phenomena into four systemic deficiencies of token counting as a measurement system, and analyzes their consequences.

4.1 Deficiency I: The Measurand Lacks Object Intrinsicness and Verifiability

The first principle of metrology is to specify the measurand before measuring. In token metering, this principle fails at three levels:

The operational definition exists but depends on the instrument. The token count can be operationally defined as "the length of the output sequence produced by a given tokenizer under a given vocabulary, normalization, and special-token rule." This definition is itself unobjectionable; the problem is that it smuggles the measurement object from "the text" into the joint product "text × tokenizer"—the measurand becomes a function of the instrument rather than a property of the object. The 40- versus 141-token discrepancy for the Chinese segment in Section 3.2 is a direct manifestation of this "definitional drift" of one and the same object under different instruments.

The measurement procedure is not published. Vendors often do not fully disclose the vocabulary version, normalization rules, and special-token treatment, rendering the definition unverifiable in practice. A user receiving a token count cannot determine under which definition it was measured.
The definition lacks object intrinsicness. Consequently, any statement about "the token count of a text" implicitly carries unstated measurement conditions; when the conditions differ, the numbers differ, and the statement itself offers no means of explaining the difference. This is the most elementary error form in metrology: measuring first and defining afterward.

4.2 Deficiency II: Incommensurable Units

Metrology requires that measurement results be expressed in units that are either identical or mutually convertible. The token satisfies neither requirement: tokens of different models are neither identical nor convertible by any public, stable ratio—because the conversion ratio depends on the text content (Section 3.2) rather than being a constant. Therefore:
Not additive: in a mixed multi-model call, the sum "this conversation consumed 100,000 tokens" has no physical meaning.
Not comparable: cross-model comparison of "price per million tokens" is, in essence, comparing a vendor quoting in feet with one quoting in meters, unless a conversion under a common benchmark is supplied.
No common benchmark: byte count could serve as a common benchmark (UTF-8 bytes are a canonical encoding of text, independent of any tokenizer), but the current ecosystem rarely quotes in bytes, forcing users to make economic decisions in incommensurable units.
Historically, this state corresponds to the metrological chaos that preceded the establishment of unit systems: every locality had its own "foot," and no two feet agreed. The significance of the SI is not that it fixed one particular foot, but that it made every foot traceable to a common standard [1].

4.3 Deficiency III: Absence of Calibration and Traceability

In metrology, the credibility of a measurement rests on a traceability chain—working instrument → calibration → reference standard → (step by step) SI unit—with the uncertainty recorded at every level. By contrast, token metering possesses no traceability chain of any kind:
No reference standard: there is no recognized standard reference text—a corpus spanning languages and domains—against which vendors' counting methods can be calibrated.
No calibration procedure: vendors self-report counts; users have no independent means of verification, and the consistency between local open-source tokenizers and server-side counts cannot be checked (Section 3).
No audit mechanism: bills present only the final number without a segmentation breakdown; even with a breakdown, the user lacks an authoritative benchmark against which to judge its correctness.
The result is a metrologically rare "vendor-unilateral metering" arrangement: the unit, the procedure, and the result are all defined and issued by one party to the transaction, while the other party can only accept them. This would be unacceptable in any mature metrological domain (electricity meters, water meters, fuel dispensers), which is precisely why legal metrology exists. It must be noted cautiously that these physical instruments are regulated under legal metrology [9], whereas software counting differs materially from physical instruments in legal object and verification; the analogy here is therefore illustrative, and the introduction of legal metrology should be regarded as a long-term direction requiring dedicated study (Section 5.2).

4.4 Deficiency IV: Inestimable Measurement Uncertainty

The GUM requires measurement results to be accompanied by an uncertainty [10]. Token counting presents a peculiar case:
Zero random dispersion. Tokenization is a deterministic algorithm; the same input under the same settings reproduces the same count. This creates an illusion of "precision," but it reflects only algorithmic determinism, not the credibility of the measurement system—the credibility problem lies precisely at the level of definition and procedure.
Divergent definitional uncertainty. The VIM defines definitional uncertainty as the component of measurement uncertainty arising from the finite amount of detail in the definition of the measurand [9]; in GUM terms this corresponds to the uncertainty component introduced by an incomplete definition of the measurand [10]. Once the definition changes (a different tokenizer, normalization, or special-token rule), the token count varies without bound—by orders of magnitude for extreme texts. This "change the ruler and the number changes" state is, in metrological terms, an unacceptable definitional uncertainty.
Unbounded systematic bias. The multilingual and domain disparity (Section 3.2) is a content-dependent systematic error that can neither be expressed by a single correction coefficient nor assessed by the user. The "4 characters ≈ 1 token" rule that holds for English prose fails for Chinese, Japanese, code, and mathematics, and the user does not know in advance where their text falls in the error distribution.
Consequently, every token-based bill should carry an uncertainty, yet under the current system that uncertainty is neither computable nor reportable. For scholarly measurements (perplexity, FLOPs/token), this means that the confidence interval of cross-model comparisons is systematically overestimated.

4.5 Implications for Billing Fairness and Cost Comparability

Taken together, the four deficiencies have three economic consequences:
The billing unit correlates weakly with cost and value. Token efficiency differs by several times across texts, so that "charging by token" is effectively "charging by a vendor-defined length" that bears no stable ratio to the information the user receives (e.g., Shannon information or semantic content). Languages with higher compression—such as Chinese—pay more per unit of information, constituting a differentiated pricing arrangement [17-18, 20-21].
Price comparison fails. Because units are incommensurable and conventions diverge (input/output, cache, hidden tokens), "price per million tokens" cannot guide cross-model cost decisions; the actual cost depends on the user's own request structure, which in turn interacts with the tokenizer—producing a metrological blind spot that users cannot cheaply resolve.
Incentives are misaligned. Because the unit is defined by the seller, the seller has an intrinsic incentive to enlarge the billing unit (for instance, by choosing a less efficient tokenization or charging for auxiliary content); absent audit and traceability, this incentive is unchecked.

4.6 Implications for Scientific Reproducibility

In scholarly settings, the deficiencies of token metering produce systematic distortion as well: a benchmark reporting "input of N tokens" is irreproducible unless the tokenizer and its version are stated; perplexity, normalized by token, loses meaning when compared across tokenizers; and cross-model comparisons of training scale ("X trillion tokens") and FLOPs/token rest on the implicit assumption that tokens are commensurable—an assumption that does not hold [14-15]. The recent rise of multilingual token-cost studies is, in essence, the research community performing, on behalf of the measurement system, the "calibration work" that the system itself should have provided—measuring, one by one, how each vendor's tokenizer deviates across languages—which in turn reveals the absence of basic metrological infrastructure [17-21].

5 TOWARD STANDARDIZATION

The analysis of Section 4 points to a general conclusion: token metering must move from "vendor-defined units" toward a measurement system that is *definable, reproducible, calibratable, and auditable.* Table 4 maps the deficiencies to their impacts and countermeasures, which are then elaborated at three levels.

Table 4 Mapping from Deficiencies to Impacts and Countermeasures

Systemic deficiency	Main impact	Countermeasure
I: measurand lacks intrinsicness	definitional drift, unverifiable	publish counting norms (§5.2); report conditions (§5.3)
II: incommensurable units	failed comparison, non-additivity	common benchmark (§5.1); standardized conversion (§5.1)
III: no calibration/traceability	vendor-unilateral metering, no audit	reference corpus, counting API, bill audit (§5.2)
IV: inestimable uncertainty	unquantifiable billing bias	report conditions (§5.3); benchmark ratios (§5.1)

5.1 Technical Level: Establishing a Common Benchmark

Against incommensurable units (Deficiency II), three candidate common benchmarks exist:
Byte/character benchmark: the UTF-8 byte count is independent of any tokenizer and is a canonical encoding of the text, hence inherently convertible across models. Billing and capacity in bytes would directly eliminate tokenizer differences, at the cost of decoupling from the model's actual computation (which is closer to token count). Baidu Qianfan's early character-based billing can be seen as a transitional form toward this benchmark [22].
Information-entropy benchmark: billing by Shannon information would charge for the quantity of information rather than its encoding [23]. This is conceptually closest to the metrological notion of a "quantity," but entropy depends on a probabilistic model and is difficult to estimate reliably for a single text; it should be treated as a long-term direction.
Standardized token conversion: while retaining token billing, vendors could publish "token-per-byte ratios under a standard reference corpus," allowing users to convert and compare on a common benchmark. This is the cheapest transitional option, but as Section 3.2 shows, a single ratio cannot cover content-dependent bias and remains only an approximation.

5.2 Metrological Level: Norms and Traceability Chains

Against the absence of definition and traceability (Deficiencies I and III), four measures can be borrowed from mature metrological systems:
Publish counting norms. Vendors should disclose the complete technical specification of their tokenizer (algorithm, vocabulary version, normalization rules, special-token treatment), making the "definition of the measurand" a verifiable public document. Open-source tokenizers (tiktoken [5], tokenizers [24]) are a good starting point, but they must be version-bound to, and consistency-tested against, the server-side implementation.
Establish a standard reference corpus. A neutral body (the scholarly community or a standards organization) should construct a reference corpus spanning languages and domains—including code and mathematics—and publish the

benchmark counts of every vendor's tokenizer on it, serving as standard reference data and the material carrier of calibration.

Provide a counting API and itemized bills. Vendors should provide an official counting endpoint, and bills should itemize input/output/cache-hit/hidden tokens so that users can reconcile and audit; third-party auditors could then compare bills against the counting API, forming a minimal external check.

Introduce legal-metrology thinking (long term). In billing contexts, token counting is a form of trade metering; following the precedent of electricity meters and fuel dispensers, the feasibility of type approval and periodic verification of counting systems should be studied. As noted in Section 4, the legal object of software counting differs from physical instruments, so this remains a long-term direction requiring dedicated research.

5.3 Scholarly-Reporting Level: Report the Measurement Conditions

Against inestimable uncertainty (Deficiency IV), the cheapest and fastest improvement is the normalization of scholarly and engineering reporting: whenever a token count is reported (input length, corpus size, FLOPs/token, or the perplexity denominator), the tokenizer name, vocabulary version, normalization, and special-token rules must be reported alongside it; cross-model comparisons must be converted to a common benchmark (e.g., bytes) or accompanied by a benchmark ratio. This requirement is exactly isomorphic to the metrological convention that "a measurement result must state its measurement conditions and uncertainty," and can be written into benchmark and publication norms immediately.

5.4 Assessment of Existing Progress

Several positive developments have emerged in recent years: tokenizers and counting libraries have been widely open-sourced (tiktoken [5], tokenizers [24]); cache pricing and peak/off-peak pricing have disclosed part of the billing conventions [11-13]; multilingual token-cost studies have supplied the first data for calibration [17-21]; and some platforms offer counting APIs and character-based billing options [22]. On the whole, however, these developments remain at the level of "individual self-disclosure": there is no public reference standard, no third-party audit, and no cross-vendor common benchmark, and billing conventions are still decided unilaterally by the seller. The ecosystem already possesses the technical components for standardization; what is missing is the organizational framework that assembles them into a measurement system—which is precisely why this paper takes metrology as its blueprint.

6 CONCLUSION

The token is the de facto universal unit of measurement of the LLM era, but it is not yet a sound unit in the metrological sense. Through the lens of measurement science, this paper has surveyed the state of token counting, reported a first-hand benchmark across 13 tokenizers and 18 multilingual texts, and identified four systemic deficiencies—the measurand lacking object intrinsicness, incommensurable units, the absence of calibration and traceability, and inestimable uncertainty—and argued that these undermine billing fairness, cost comparability, and scientific reproducibility. Measured against the history of metrology, token metering today occupies the same "pre-SI" stage as the chaotic systems of weights and measures that preceded the SI; two centuries of metrology have already supplied the complete roadmap out of that chaos—well-defined quantities, reproducible units, stepwise traceable calibration chains, assessable uncertainties, and the standard reference data and public norms that carry them.

We recommend advancing the standardization of token metering at three levels: establishing common benchmarks such as bytes at the technical level; establishing public counting norms, a standard reference corpus, a counting API, bill auditing, and (in the long term) a legal-metrology framework at the metrological level; and mandating the reporting of measurement conditions at the scholarly level. These measures do not require abandoning token billing; they require transforming the token from a "vendor-defined count" into a "definable, reproducible, calibratable, auditable measurement result." Only then can "price per million tokens" become a reliable comparison tool, can cross-model performance and cost comparisons rest on a solid measurement basis, and can the LLM economy and LLM science move from mutual unintelligibility toward commensurable, reproducible maturity.

7 LIMITATIONS

As a conceptual survey, this paper has several limitations. First, the measurements of Section 3.2 and Appendix A cover 13 tokenizers and 18 multilingual texts, but the corpus size and language coverage remain limited, and the open-source tokenizers were obtained through a third-party mirror rather than validated against the official releases. Second, the billing-convention summary of Section 3 relies on public pricing pages and secondary reports; vendor policies are time-sensitive and may have changed after writing, so readers should defer to the latest official releases. Third, we did not provide a quantitative feasibility analysis of the alternative benchmarks (byte-based, entropy-based). Fourth, the feasibility and legal basis of the legal-metrology analogy require dedicated legal research. These limitations also point to future work: larger cross-tokenizer benchmarks, continued tracking of billing conventions, and quantitative evaluation of alternative billing units.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This study was supported by the Science and Technology Planning Project of Maoming City, Guangdong Province (Grant No. 2025036).

REFERENCES

[1] BIPM. The International System of Units (SI Brochure), 9th edition, 2019.
[2] Mielke S J, Alyafeai Z, Salesky E, et al. Between Words and Characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. Transactions of the Association for Computational Linguistics, 2021, 9: 229–246.
[3] Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of ACL, 2016.
[4] Radford A, Wu J, Child R, et al. Language Models are Unsupervised Multitask Learners. OpenAI Technical Report, 2019.
[5] OpenAI. tiktoken: a fast BPE tokeniser for use with OpenAI's models. 2026. https://github.com/openai/tiktoken; OpenAI Platform documentation (tokenizer and billing). https://platform.openai.com/docs.
[6] Wu Y, Schuster M, Chen Z, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144, 2016.
[7] Kudo T, Richardson J. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In: Proceedings of EMNLP, 2018.
[8] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971, 2023.
[9] JCGM 200:2012. International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM). BIPM, 2012.
[10] JCGM 100:2008. Evaluation of Measurement Data—Guide to the Expression of Uncertainty in Measurement (GUM). BIPM, 2008.
[11] OpenAI. API Pricing and cached-input pricing documentation. https://openai.com/api/pricing/.
[12] DeepSeek. API pricing and cache billing documentation (including the August 2026 peak/off-peak adjustment). https://api-docs.deepseek.com/quick_start/pricing.
[13] Anthropic. Prompt Caching documentation. https://docs.anthropic.com/en/docs/build-with-claude/prompt-caching.
[14] Kaplan J, McCandlish S, Henighan T, et al. Scaling Laws for Neural Language Models. arXiv:2001.08361, 2020.
[15] Hoffmann J, Borgeaud S, Mensch A, et al. Training Compute-Optimal Large Language Models. arXiv:2203.15556, 2022.
[16] DeepSeek-AI. DeepSeek-V3 Technical Report. arXiv:2412.19437, 2024.
[17] Ahia O, Kumar S, Hila Gonen, et al. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. In: Proceedings of EMNLP, arXiv:2305.13707, 2023.
[18] Deng N, Solatorio AV, Vicente GS, et al. Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-economic Disparities and Reduced Utility for Non-English Speakers. arXiv:2410.10665, 2024.
[19] Ali M, Fromm M, Thellmann K, et al. Tokenizer Choice For LLM Training: Negligible or Crucial? arXiv:2310.08754, 2023.
[20] Lundin JM, Zhang A, Karim N, et al. The Token Tax: Systematic Bias in Multilingual Tokenization. arXiv:2509.05486, 2025.
[21] Teklehaymanot H K, Nejdl W Tokenization Disparities as Infrastructure Bias: How Subword Systems Create Inequities in LLM Access and Efficiency. arXiv:2510.12389, 2025.
[22] Baidu AI Cloud. Qianfan platform billing documentation and ERNIE token calculation notes. 2026. https://cloud.baidu.com/doc/qianfan/.
[23] Shannon C E. A Mathematical Theory of Communication. Bell System Technical Journal, 1948, 27(3): 379–423.
[24] HuggingFace. tokenizers: Fast State-of-the-Art Tokenizers. 2026. https://github.com/huggingface/tokenizers.

APPENDIX A: A MULTILINGUAL TOKEN METERING BENCHMARK

To extend the illustrative measurement of Section 3.2 into a reproducible systematic benchmark, we constructed a multilingual × multi-domain token metering benchmark. All measurements were performed with tiktoken 0.13.0 and tokenizers 0.21.4 using publicly released tokenizers; the benchmark data and reproduction scripts are available from the corresponding author upon request.

A.1 Benchmark Design

Tokenizers: 13 in total, covering three OpenAI generations (gpt2 / cl100k_base / o200k_base) and ten open-source model tokenizers, with vocabulary sizes from roughly 30,000 to 200,000 (Table A1).

Corpus: 18 segments, comprising the parallel texts of Article 1 of the Universal Declaration of Human Rights (UDHR) in 12 languages, plus technical descriptions (English and Chinese), code (Python/JavaScript), mathematical formulas, and special symbols.
Metrics: token count, UTF-8 byte count, character count, token/byte, token/character; cross-tokenizer statistics include minimum, maximum, ratio, and coefficient of variation.

Table A1 Tokenizers Used in the Benchmark and Their Vocabulary Sizes

Tokenizer	Vocab size	Source
LLaMA-2	32,000	NousResearch mirror
Mistral-7B	32,000	mistralai
Phi-3	32,011	microsoft
GPT-2 (gpt2)	50,257	tiktoken
Yi-6B	63,992	01-ai
InternLM2	92,544	internlm
GPT-3.5/4 (cl100k_base)	100,277	tiktoken
Baichuan2	125,696	baichuan-inc
LLaMA-3	128,256	NousResearch mirror
DeepSeek-V3	128,815	deepseek-ai
GLM-4	151,329	zai-org
Qwen2.5	151,665	Qwen
GPT-4o (o200k_base)	200,019	tiktoken

A.2 Multilingual Token Inflation

With English as the baseline, Table A2 reports the mean ratio of token counts over the 13 tokenizers for each language, together with the cross-tokenizer range. Figure 1 presents these mean ratios graphically.

Table A2 Token Inflation Relative to English (UDHR Parallel Texts)

Language	Mean ratio	Cross-tokenizer range
English	1.00	—
Chinese	1.17	0.35–2.48
German	1.32	0.97–1.82
Spanish	1.32	1.01–1.76
French	1.48	1.13–1.88
Russian	1.89	1.12–5.21
Japanese	2.17	0.99–3.55
Arabic	2.29	1.30–4.21
Korean	2.69	1.11–5.97
Vietnamese	2.72	1.44–5.45
Thai	3.78	1.58–10.52
Hindi	5.04	1.61–12.15

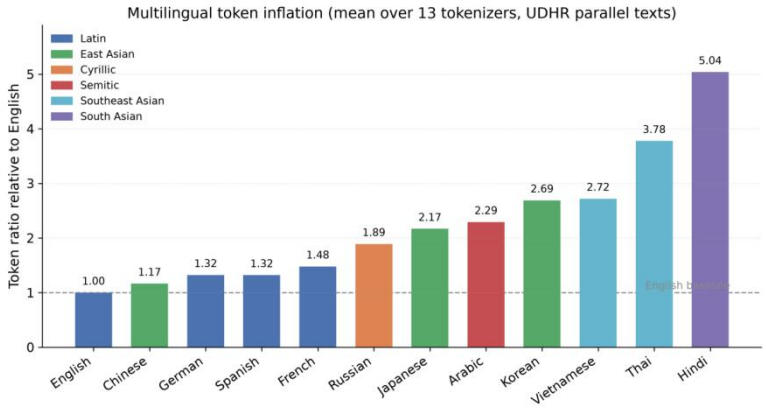


Figure 1 Multilingual Token Inflation Relative to English (Mean over 13 Tokenizers)

A.3 Chinese-Segment Comparison across Tokenizers

Table A3 reports the token counts of the same 43-character Chinese text across tokenizers. Figure 2 presents the same comparison, with models grouped as Chinese-oriented or Western-oriented.

Table A3 Token Counts of the Same 43-Character Chinese Text under Different Tokenizers

Tokenizer	Tokens	token/char

Tokenizer	Tokens	token/char
DeepSeek-V3	27	0.63
Baichuan2	27	0.63
GLM-4	28	0.65
Yi-6B	28	0.65
Qwen2.5	29	0.67
InternLM2	30	0.70
GPT-4o (o200k_base)	37	0.86
LLaMA-3	39	0.91
Mistral-7B	45	1.05
GPT-3.5/4 (cl100k_base)	52	1.21
Phi-3	62	1.44
LLaMA-2	63	1.47
GPT-2 (gpt2)	82	1.91

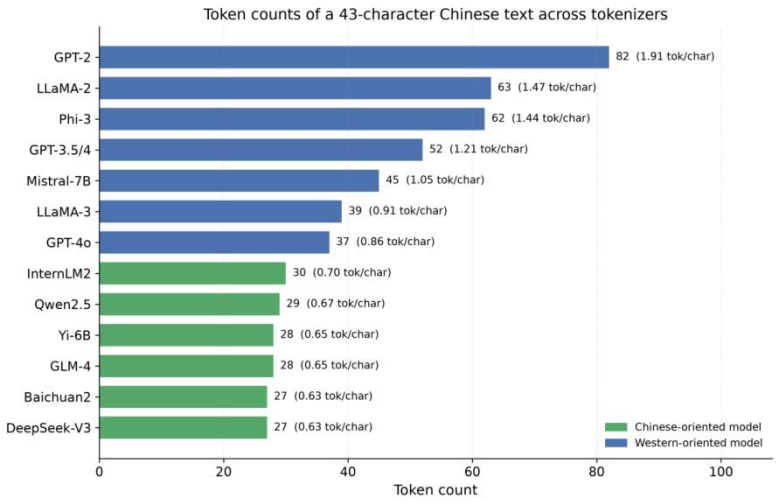


Figure 2 Token Counts of a Chinese Segment across Tokenizers

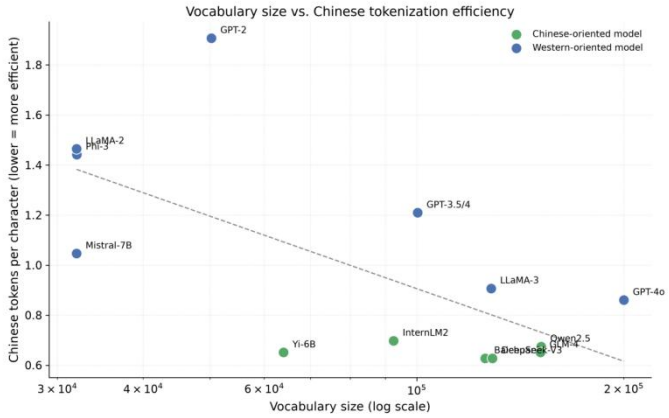


Figure 3 Vocabulary Size versus Chinese Segmentation Efficiency

A.4 Findings

The benchmark reveals four regularities. First, token inflation grows with the distance between the writing system and the Latin alphabet (Hindi 5.04×, Thai 3.78×, versus only about 1.3–1.5× for Western European languages). Second, the efficiency of the same language varies by a factor of 7–12 across tokenizers (Hindi 1.61–12.15×), showing that "how many tokens a language consumes" fluctuates sharply with the tokenizer. Third, Chinese-oriented models (Baichuan2, DeepSeek-V3, GLM-4, Qwen2.5, Yi-6B, InternLM2) segment Chinese substantially more efficiently than Western-oriented models (Phi-3, LLaMA-2, GPT-2), with the token count of the same 43-character text ranging from 27 to 82—approximately a threefold difference (Table A3, Figure 2). Fourth, Chinese segmentation efficiency is positively but not decisively correlated with vocabulary size (Figure 3): among models with vocabulary sizes of approximately 150,000, GLM-4 and Qwen2.5 segment Chinese efficiently, whereas o200k, despite a larger vocabulary, performs worse than most Chinese-oriented models on Chinese text—indicating that the language composition of the training corpus matters more than vocabulary size.