universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

The Geography of Token Efficiency in Large Language Models

verfasst von | submitted by

Florian Mathias Venier

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 856
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Kartographie und Geoinformation
Betreut von | Supervisor:	Univ.-Prof. Dr. Krzysztof Janowicz

Abstract

Tokenization is a core preprocessing step in large language models and directly influences how text is represented and processed. However, token efficiency is typically examined as a language- or model-specific property, without considering its spatial distribution. This thesis addresses this gap by analyzing token efficiency as a global spatial phenomenon.

Using a global H3 grid and OpenStreetMap-based place names, token efficiency is measured and compared across three contemporary tokenizer architectures. The analysis combines descriptive statistics, spatial autocorrelation, clustering methods, and cross-model comparison to examine how efficiency varies across regions and between models. In addition, token efficiency is related to external data layers, including population density, dominant writing system, carbon intensity of electricity, and internet penetration.

The results show that token efficiency exhibits clear spatial patterns, forming coherent regions of relatively high and low efficiency that remain stable across different spatial aggregations. While the analyzed tokenizers differ in their overall efficiency levels, their spatial distributions are highly similar. This indicates that the choice of the model affects the degree of efficiency, but not its geographic structure

These findings suggest that spatial variation in token efficiency is shaped by broader structural factors rather than individual model design. By introducing a spatial perspective on tokenization, this thesis highlights how differences in linguistic representation are unevenly distributed across geographic space and may have implications for computational cost and access to AI systems.

Kurzfassung

Tokenisierung ist ein zentraler Vorverarbeitungsschritt in großen Sprachmodellen und beeinflusst direkt, wie Text repräsentiert und verarbeitet wird. Dennoch wird Token-Effizienz meist als sprach- oder modellspezifische Eigenschaft betrachtet, ohne ihre räumliche Verteilung zu berücksichtigen. Diese Arbeit greift diese Forschungslücke auf, indem sie Token-Effizienz als globales räumliches Phänomen analysiert.

Auf Basis eines globalen H3-Rasters und von OpenStreetMap abgeleiteten Ortsnamen wird die Token-Effizienz gemessen und über drei aktuelle Tokenizer-Architekturen hinweg verglichen. Die Analyse kombiniert deskriptive Statistik, räumliche Autokorrelation, Clusterverfahren sowie Modellvergleiche, um zu untersuchen, wie sich Effizienz räumlich verteilt und zwischen Modellen unterscheidet. Zusätzlich wird die Token-Effizienz mit externen Datensätzen in Beziehung gesetzt, darunter Bevölkerungsdichte, dominantes Schriftsystem, CO -Intensität der Stromerzeugung und Internetdurchdringung.

Die Ergebnisse zeigen, dass Token-Effizienz klare räumliche Muster aufweist und sich in zusammenhängenden Regionen mit relativ hoher und niedriger Effizienz organisiert, die über verschiedene räumliche Aggregationen hinweg stabil bleiben. Obwohl sich die untersuchten Tokenizer in ihrem allgemeinen Effizienzniveau unterscheiden, sind ihre räumlichen Verteilungen sehr ähnlich. Dies deutet darauf hin, dass die Modellwahl das Ausmaß der Effizienz beeinflusst, nicht jedoch deren geografische Struktur.

Die Ergebnisse legen nahe, dass räumliche Unterschiede in der Token-Effizienz eher durch übergeordnete strukturelle Faktoren als durch individuelles Modelldesign geprägt sind. Durch die Einführung einer räumlichen Perspektive auf Tokenisierung zeigt diese Arbeit, wie Unterschiede in der sprachlichen Repräsentation ungleich über den geografischen Raum verteilt sind und welche Implikationen sich daraus für Rechenaufwand und den Zugang zu KI-Systemen ergeben können.

Acknowledgment

This thesis would not have been possible without the support of many people.

I would like to express my sincere gratitude to my supervisor Krzysztof Janowicz for his guidance, patience, and detailed feedback throughout the entire process. His input shaped this thesis in ways I could not have managed on my own.

I would also like to thank my friends Bianca, Marcel and Bo, who read through drafts and helped me whenever my writing was unclear. Their honest feedback improved the quality of this work considerably.

My deepest thanks go to my family. To my father Wolfgang, my mother Sabine, and my brothers Fabian and Lukas, thank you for your constant support and encouragement throughout my studies. You were there through all the difficult times and never stopped believing in me. By giving me the financial support to study without worry, you made all of this possible in the first place. I could not have done this without you.

Finally, I would like to thank my employer Wiener Linien for offering me a new position before I had even completed my degree. That trust meant a great deal to me and motivated me to bring this thesis to a close.

Contents

List of Figures

List of Tables

1 Introduction

Large language models are being used globally to process text across a wide range of languages and geographic contexts. A key step in this process is *tokenization*, in which raw text is segmented into discrete units before model computation. Tokenization affects how efficiently text is represented but also influences computational cost and model behavior. Previous research has shown that tokenization efficiency varies across languages and that this variation is systematic and not incidental. (Sennrich et al., 2016)

Follow-up studies have compared different tokenization approaches and have demonstrated that there are efficiency differences between tokenizer architectures and multilingual settings. When comparing BPE, Unigram, and WordPiece tokenizers, they found that efficiency varies across linguistic contexts, with some languages consistently encoding less compactly than others. (Rahman et al., 2024; Velayuthan & Sarveswaran, 2025) At the same time, recent work has argued that evaluating tokenizers purely based on global averages or benchmark-level metrics can overlook important distributional effects. (Schmidt et al., 2024). While these studies have advanced our understanding of token efficiency, they have primarily examined efficiency at the level of individual languages or benchmark datasets. These efficiency differences have practical consequences that go beyond technical performance. Commercial large language models charge users per token processed. This means that a user working in a language that requires more tokens to express the same information pays more for the same task. (Ahia et al., 2023) Beyond cost, higher token counts also mean slower processing and greater computational demand.

The systems we build reflect the conditions under which they were built. Large language models trained predominantly on English or Western text encode structural biases that disadvantage other languages and communities. (Bender et al., 2021). Most NLP research focuses on a small number of high-resource languages, leaving the majority of the world´s linguistic diversity underrepresented in both training data and model evaluation. (Joshi et al., 2020) This is not simply a coincidence or an oversight, but reflects broader structural patterns in how AI systems are developed and deployed. The question is not only which languages are well represented, but where these advantages and disadvantages are located and if they form spatial patterns that persist across systems.

Geography has a clear role in how digital systems distribute their benefits and disadvantages. Digital tools that appear to be global and neutral in design consistently produce results that are shaped by the geography of data production, favoring regions with strong digital infrastructure and high representation in training corpora. (Ballatore et al., 2017) The development and deployment of AI systems follow the same logic. Access, infrastructure and the representation of training corpora are unevenly distributed across the world, and these distributions mirror existing global inequalities rather than overcoming them. (Peng, 2024; Shi et al., 2025) When AI-generated content is examined for geographic diversity, models are found to consistently favor regions with rich digital data while underrepresenting large parts of the world. (Liu et al., 2025). For example Goodchild (2007) and Janowicz (2023) both argue that AI does not exist outside of the world it was build in and that its effects are always spatially situated. Biases do not only emerge at the level of model output, but also at earlier stages such as how information is segmented and represented. Tokenization is one of these earlier stages, and if its efficiency is uneven across languages, and languages are unevenly distributed across space, then tokenization itself becomes a mechanism through which spatial inequality enters large language models before any higher level processing takes place.

An area that remains largely under-explored is the question of whether differences in token efficiency show a spatial structure. Languages, their naming conventions, and textual data are not randomly distributed across the world. They are embedded in geographical contexts that are shaped by history, administration, and culture. If token efficiency varies systematically across languages and languages are distributed across space, then the efficiency of tokenization itself may form coherent geographic patterns. This connection has not yet been made in existing research. Studies of tokenization remain focused on individual languages or model architectures, while GeoAI and AI ethics research have not considered tokenization as a spatial variable. This is the gap this thesis wants to address.

Place names and geographical terms provide a useful approach to answering this question. They are a globally distributed and linguistically diverse form of text that reflects the spatial organization of language. By investigating token efficiency through a spatial view, this thesis offers a way to examine if these differences form coherent geographic patterns or appear as isolated, language specific effects.

This question is also important when discussing digital inequality and sustainability. Critical research has shown that LLMs reflect, and reproduce uneven distributions of linguistic and cultural representation. (Bender et al., 2021; Joshi et al., 2020). Related work on the environmental implications of LLMs suggests that computational cost is closely tied to efficiency choices and that efficiency gains are not always evenly distributed among

users or regions (Shi et al., 2025). By identifying where tokenization inefficiencies lie, this thesis aims to provide the necessary work for future research on linguistic based energy disparities and encourages a closer examination of where inefficiencies are located and whether they exist across model architectures.

This thesis addresses these questions by analyzing token efficiency as a spatial concept. Using a global H3 grid and OpenStreetMap (OSM)-extracted place name based textual inputs, it measures token efficiency across spatial cells worldwide, aggregating the results for macro regions and countries. Three different tokenizer architectures are compared to investigate whether spatial patterns are dependent on the design choices of the models or reflect more general structural properties of the training data being tokenized. To situate these patterns within broader geographic and environmental contexts, token efficiency is further compared against four external data layers: population density, dominant writing system, carbon intensity of electricity and internet penetration rate.

Hence, the research questions of this thesis are as follows:

RQ1: How much does token efficiency vary worldwide across different regions?

RQ2: How do different tokenizers perform globally in terms of token efficiency?

RQ3: How are the observed patterns of token efficiency related to broader questions of digital inequality and sustainability?

This thesis extends existing research beyond language- and dataset centered analyses by adopting a spatial perspective on token efficiency. It aims to provide empirical evidence on how efficiency differences are distributed across space and how these differences exist across tokenizer architectures, offering a new perspective on how tokenization efficiency is structured across the globe. By making these patterns visible and measurable, this thesis establishes token efficiency as a spatial baseline for evaluating fairness in large language models. It shows that tokenization, typically treated as a neutral preprocessing step, actively shapes how languages and places are represented within these systems before any higher level processing takes place. The findings contribute to ongoing discussions in AI ethics and GeoAI by providing a concrete, and reproducible method for identifying where tokenization disadvantages are located and how stable they are across independently developed systems.

2 Related Work and Theoretical Background

Large Language Models (LLMs) have become a central component in modern artificial intelligence. They are trained from textual data in many different languages and are able to generate, summarize, interpret, and translate information with growing proficiency. At the heart of this capability is the tokenization process, which determines how an LLM reads in data. Tokenization breaks data into discrete units so that they can serve as input for neural architectures. The way this process is designed affects how efficient LLMs are for different languages. Prior research shows that tokenization efficiency varies systematically between languages. Petrov et al. (2023) show that cross-lingual tokenization leads to consistent differences in segmentation patterns between morphologically simple and morphologically complex languages. Similarly, Rahman et al. (2024) demonstrate that standard tokenizers produce inflated token counts for certain linguistic structures and scripts, resulting in measurable differences in computational cost and representation quality across languages.

Tokenization is an important factor in determining how well languages align with the computational structure of an LLM. Languages, like English, that have simple word boundaries, are represented with few tokens. Morphologically complex languages, like Turkish or Finnish, are split into many small parts, which increases the number of tokens per sentence. This is in direct relation to cost, speed, and environmental impact. Therefore, tokenization not only prepares text for computation but also potentially introduces inequality. (Mor, 2025; Petrov et al., 2023) Mor (2025) and Santorelli et al. (2024) explain that technological participation increasingly depends on how compatible a language is in such regard. In this sense, tokenization connects digital inclusion to linguistic structure.

While much of the research on this topic has focused on the architecture and training of LLMs, the linguistic preprocessing step of tokenization shapes how text enters a model and how meaning is processed through the system. Janowicz (2023) notes that GeoAI spatial computing depends on the representation of information in its smallest units. This logic also applies to LLMs, as biases begin in the way data are discretized. Once a language is split into inefficient tokens, disadvantages may propagate through the following stages,

including embedding and final inference. This means that tokenization reflects social and linguistic hierarchies by determining which languages are processed efficiently and which are not.(Bender et al., 2021; Mielke et al., 2021)

From a geographic perspective, tokenization also has spatial implications, as languages are not distributed evenly across the globe. Differences in token efficiency can be seen as digital inequalities with geographic patterns. Some languages are concentrated in regions with strong digital infrastructure and high representation in the training corpora, while others are predominantly spoken in areas with limited technological visibility. Ballatore et al. (2017) shows that even search engines, often seen as global and neutral, produce results that are shaped by geography and favor certain regions over others. Their study shows that descriptions of many regions, particularly in parts of Africa and Asia, are frequently authored and framed from Anglophone and Western perspectives rather than from within those regions themselves. This demonstrates how digital representation is shaped by centers of data production. Applying this logic to AI suggests that the linguistic and computational infrastructure may reproduce existing imbalances in representation rather than minimizing them. Shi et al. (2025) and Peng (2024) argue in a similar way and explain that the geography of AI mirrors global patterns of inequality. By mapping token efficiency, these asymmetries should be made visible. Tokenization differences are not isolated linguistic anomalies, but rather, they are spatially structured imbalances, that influence who benefits most from AI systems. Given that linguistic and digital infrastructures are unevenly distributed across the globe, persistent differences in token efficiency can contribute to broader patterns of computational and economic inequality. (Ballatore et al., 2017; Peng, 2024; Shi et al., 2025)

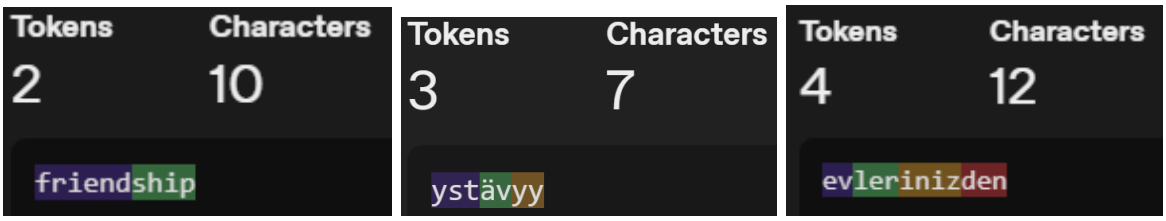
Tokenization efficiency is directly tied to environmental sustainability. Every token processed by a model requires a physical expenditure of energy and water for cooling data centers (Li et al., 2023). Recent research by Shi et al. (2025) and Petrov et al. (2023) shows that this resource use is not distributed equally. Because many standard tokenizers are optimized for English, they often require significantly more tokens to represent the same idea in other languages. This means that users in different regions effectively pay a higher environmental and financial price to access the same information (Ali et al., 2024). Consequently, the ecological footprint of AI is not just a technical issue, but one that scales based on language and geography (Petrov et al., 2023; Shi et al., 2025).

2.1 Tokenization and LLMs

Tokenization co-determines how large language models read and represent language. Essentially, they divide data such as written text or source code into small chunks ready

for ingestion for the vector-based input layer of neural networks. The three most used algorithms are Byte Pair Encoding (BPE), Unigram, and WordPiece. Each of these algorithms split text into units based on statistical patterns identified in large training corpora. These units may correspond to full words, frequent subwords, syllables, or even individual characters, depending on the tokenizer design and vocabulary construction. Tokenizers aim to maintain a manageable vocabulary and, at the same time, preserve as much linguistic meaning as possible. However, their performance varies across languages, as words and characters do not always follow similar patterns. Some tokenizers recognize frequent character combinations, while others work with entire syllables or morphemes (Sennrich et al., 2016; Velayuthan & Sarveswaran, 2025).

Tokenization plays a key role in how well language fits into a model‘s structure. A tokenizer that is primarily trained using English is able to efficiently compress sentences but struggles with languages with complex morphology, as it tends to split its words into many parts. This increases the number of tokens that are required for the same meaning. For example, the English word “friendship” is often encoded as two subwords: “friend” and “ship”; see Fig. 2.1a.



(a) English (b) Finnish (c) Turkish

Figure 2.1: Token counts across languages

while the Finnish equivalent, ”ystävyy” (“in a friendship”) has to be divided into several smaller tokens; see Fig. 2.1b. Turkish shows a similar pattern: evlerinizden (”from your houses”) becomes ev, ler, iniz, den; see Fig. 2.1c.

These longer token chains demonstrate that morphologically rich languages produce more tokens to express the same content, which increases their computational cost. (Rahman et al., 2024) This pattern appears systematically across multilingual models, making some languages more expensive to process, as Petrov et al. (2023)demonstrates. Since each additional token increases the number of computational operations required during inference, higher token counts are often linked to greater energy consumption and water usage in large language models. (Shi et al., 2025). That means that tokenization inefficiency not only increases monetary cost, but also contributes to the environmental footprint of AI systems.

The way tokens are created affects both the quality of the model and the speed. According

to Goldman et al. (2024), efficient tokenization, which represents text with as few tokens as possible, improves both accuracy and speed. Algorithms handle these shorter sequences more effectively and also require less computation to achieve the same level of quality. On the other hand, inefficient tokenization lengthens inputs and increases processing time. These differences reveal how small design choices influence an algorithm´s linguistic capacity and usability for different languages (Goldman et al., 2024).

Tokenization is not only a technical mechanism, implicitly it is also an ethical and cultural one. Ali et al. (2024) argues that the design choices made at this stage ultimately determine which languages are viewed as standard and which are viewed as exceptions. When Western languages are the main source of training, they define efficiency based on their linguistic structure. Mor (2025) adds that linguistic access has become a requirement to participate in digital environments. That means that fairness in AI begins with the ability to represent all languages within these systems (Ali et al., 2024; Bender et al., 2021; Mor, 2025).

When discussing AI fairness, we usually mean the absence of systematic biases that disadvantage certain groups. Mehrabi et al. (2021) describe this as unfair discrimination caused by the data itself, the models´ design, fine tuning, or a combination of the these. Bias may enter at early stages, e.g., during data collection or reprocessing, long before a model generates a response. This suggests that fairness is not just a final characteristic of a model, but the cumulative result of every design choice made along the way. For language models, this means that fairness begins with how text is represented. If a model requires additional tokens to express the same idea in one language compared to another, that language faces a literal tax in terms of cost and processing speed. While these disparities might be invisible to the end-user, they may create an unequal playing field from the start.

The tokenization practices created for culturally dominant languages influence the spatial and cultural impact of AI. As already discussed, algorithms primarily trained on English perform best in regions where this language is dominant. On the other hand, languages with complex morphology and sparsely used in training data often lead to reduced model performance. These differences are not random. Instead they may follow the geography of digital infrastructure and data availability. Shi et al. (2025) and Peng (2024) point out that the geography of AI reinforces historic forms of global inequality as digital power is accumulated in regions with the highest linguistic compatibility. Shahid et al. (2025) describe this as a form of computational colonialism, where the languages of the Global North dominate the training material and, thus, the digital landscape. Token inefficiency can be regarded as a quantifier for this dominance. Regions that speak morphologically complex or low-resource languages face slower, more expensive, and less accurate model

performance. That means that AI can make existing centers of influence even stronger and gives advantages to regions that are already well represented in global datasets (Peng, 2024; Shahid et al., 2025; Shi et al., 2025).

From this perspective, each token symbolizes a design choice on what counts as a meaningful unit of text and what does not. Bender et al. (2021) highlight that these design choices have cultural and ethical consequences. If a language is split into a large number of small tokens, it may lose its internal logic, rhythm, and idioms. Rich linguistic systems can therefore be transformed into simplified sequences that fit computational requirements. As a result, certain forms of linguistic diversity may become harder to detect, and cultural complexities can be reduced to statistically interpreted fragments. Janowicz (2023) and Wang et al. (2025) note that this simplification mirrors wider social processes in which complex realities are reduced to data for the sake of efficiency. The more efficiently text is processed, the more likely it may become standardized, potentially reducing aspects of its original expressive richness. Also, by compressing language into uniform digital units, tokenization may contribute to embedding them within the global architecture of AI. Recognizing this pattern can help to understand how digital systems normalize language. It also helps us understand how local forms of expression risk being homogenized within the global circulation of data (Janowicz, 2023; Mielke et al., 2021; Peng, 2024; Shi et al., 2025; Wang et al., 2025).

2.2 Token Efficiency and Linguistic Diversity

Token efficiency measures how economically tokenizers encode linguistic meaning or how many tokens are required to represent a given unit of text. High token efficiency means that a model can represent information compactly. Low token efficiency means that words or morphemes are split into several smaller units. Token efficiency is influenced by linguistic structures, writing systems, and the morphological complexity of each language (Petrov et al., 2023).

Morphological typology has the strongest impact on token efficiency. Here, languages can be grouped into four categories: isolating, agglutinative, fusional, and polysynthetic. Each category has different effects on token segmentation (Velayuthan & Sarveswaran, 2025). Isolating languages like Mandarin consist mainly of single morphemes that often align well with token boundaries. An example of this would be: “我爱你” (wǒ ài nǐ) literally means “I love you”. Each word is a morpheme: wǒ (I), ài (love), and nǐ (you). In agglutinative languages like Turkish or Finnish, grammatical relations are expressed by adding many suffixes to a single stem. This provides long words that are divided into multiple tokens. An example of this would be: “kitapçıdaydım”, which means “I was at the bookstore”.

Broken down into morphemes: Kitap (book), çı (seller), da (at/in), ydı (was) and m (first person marker "I"). Fusional Languages, such as Spanish or Russian, have endings that merge several grammatical meanings. This leads to mixed efficiency, as it depends heavily on the tokenizers design. An example of this would be: "hablamos", which means "we speak" or "we spoke". The ending amos fuses person (we), number (plural) and tense (past or present) in one unit. Words from polysynthetic languages express what is often a whole sentence in other languages. A single word might include verb, noun, and object markers. An example of this is: "tusaatsiarunnanngittualuujunga", which translates to "I can't hear very well".

These linguistic differences are further amplified by writing systems. Languages that use syllabaries or logographic scripts express meaning through symbols rather than letters, creating even more challenges for subword-based tokenizers. Rahman et al. (2024) demonstrate in his work that these tokenizers, which were originally developed for alphabetic languages, excessively segment character-based scripts. This resulted in inflated token counts. Algorithms like Byte Pair Encoding and WordPiece work well for Latin scripts, but over-segment languages like Japanese and Thai. This means that token efficiency is not universal, but strongly depends on how tokenizer architecture interacts with orthographic systems. Communities using non-Latin scripts may be more likely to experience inefficiency in tokenization, which highlights an often-overlooked dimension of digital inequality (Rahman et al., 2024).

These disparities lead to practical and economic consequences. Commercial LLMs often charge the user per token processed. Meaning that the use of low-efficiency languages costs more for the same amount of information. According to Ahia et al. (2023), producing an equivalent output in Korean, Mandarin, or Finnish can often require more than twice as many total tokens as in English. This increased computational and economic cost reinforces an indirect hierarchy in which some languages enjoy faster model access than others. Token efficiency can also be understood as an economic measure and not only a technical one, as it shapes biases on who benefits most from its technologies (Ahia et al., 2023).

Researchers have started to find ways to improve token efficiency across languages. For example, Rahman et al. (2024) proposes the use of linguistically aware tokenizers that adapt segmentation rules to the writing system and morphology of each language. Rather than relying on a single universal subword vocabulary, their method creates language-dependent token lists that are based on morphology, phonology, and orthographic conventions of each language. For morphology rich languages, this means recognizing common morphological patterns and grouping them into single units rather than splitting them into different fragments. In character-based languages, the approach treats characters or

syllables as stable single tokens, which reduces over-segmentation and improves semantic consistency. This increases token efficiency, produces more stable embeddings, and improves translation accuracy in multilingual settings. However, the authors also point out some trade-offs. For example, creating these separate vocabularies for each language drastically increases storage demand and complicates model integration, since languages no longer share parameters across linguistic boundaries. Another significant drawback is an increase in training and, thereby, in the cost of commercial LLMs. Despite these challenges, their findings show that adapting tokenization to linguistic structure can significantly reduce bias and improve fairness across languages (Rahman et al., 2024).

2.3 Spatial Data and GeoAI

While tokenization studies have largely focused on linguistic and computational aspects, the geospatial impact of its effects has received minimal attention so far. Even though Janowicz (2023) reminds us that AI does not exist in a spatial vacuum. It is much rather rooted in geographical contexts that are shaped by infrastructure, language, and access. The field of Geographic Artificial Intelligence (GeoAI) integrates spatial data with machine learning to analyze how algorithms interact with the environment and humans. Applying GeoAI to language models allows researchers to visualize how and where linguistic and digital inequalities appear, making abstract computational disparities visible and mappable (Goodchild & Longley, 2021).

Recent work develops this idea further. Janowicz et al. (2025) introduce so called "geo-foundation models" (GeoFM), which aim to include geographic context directly within AI systems instead of treating it as external information. These models are intended to learn representations of geographic space across different types of data, allowing spatial patterns to be captured more consistently. This suggests that geography can play a role in the design of models, rather than being limited to the evaluation of their outputs. From this perspective, tokenization can be understood as more than a purely technical step. Differences in how languages are segmented may reflect uneven representation of geographic and linguistic contexts in the data on which models are trained.

This suggests that tokenization inefficiency can be understood not only as a linguistic problem but also a spatial problem. Because languages are distributed unevenly across the planet, they often correspond to disparities in technological infrastructure and economic development. Shi et al. (2025) claim that AI should be examined through a geographical lens, because data availability and computational access can vary considerably between regions. Linguistically diverse areas but economically developing regions, like

Sub-Saharan Africa or Southeast Asia, often experience higher rates of inefficiency, reflecting structural barriers to participation in global AI systems. Spatial analysis of token efficiency can help identify these digital "cold spots", showing the influence of geography on the inclusion of AI technologies (Shi et al., 2025).

Spatial analysis often requires discretisation. Spatial data are broken down into (uniform) units, such as cells, vectors, or pixels, with the goal of making them computable. The same logic applies to language through tokenization. Both processes simplify realities into discrete representations, enabling analysis, but also leading to biases. Together, they provide a framework for studying how representation affects accessibility. Brodsky (2018) introduces hexagonal H3 grid cells as a consistent way to aggregate information on a global scale. They also make it possible to connect spatial locations with linguistic efficiency of LLMs, acting as a bridge between geography and computation (Berrill et al., 2022; Birch et al., 2007; Brodsky, 2018; Janowicz, 2023)

From a theoretical perspective, this is directly related to research on digital inequality. Graham et al. (2015) are pointing out that access to digital infrastructure is influenced by clear spatial patterns that favor Anglophone, urban, and wealthy regions. Adding token efficiency to this framework enables a way to quantify how the performance of a language aligns with these existing hierarchies. Showing once again that linguistic factors that influence tokenization may mirror global patterns of power and capital (Graham et al., 2015).

Recent research by Liu et al. (2025) builds on the analytical potential of GeoAI, because they introduce methods to operationalize geographic diversity when evaluating AI-generated content. The study investigates the representation of different parts of the world in AI-generated content and whether spatial coverage is distributed evenly across the globe. By analyzing place names and geographical references produced by LLMs, they presented indicators that measure the extent and balance of geographical representation. The findings show that AI consistently favors regions with rich digital data and a robust economic presence, primarily North America and Western Europe, while underrepresenting much of the global south. This indicates the uneven data foundations on which these models are trained.

Liu et al. (2025) argue that geographic diversity should be treated as a key aspect of AI evaluation, not as a minor issue. The way in which spatial references are generated shows not only cultural biases but also structural dependencies on data-rich areas. It shows that fairness in AI must include the representation of place, as well as people and language. The parallels to token efficiency are clear: both concepts measure how well computational systems translate global diversity into digital form, and both measures try to identify where this fails. This is why integrating these measures helps give a complete view of

digital inequality, showing how language and space together determine who is visible and who remains overlooked. This moves the focus from abstract debates about fairness to observable evidence. Placing token efficiency within geographic space raises new questions: Where are inefficiencies most concentrated? Which linguistic regions face the highest computational and financial burden? How do these patterns align with indicators such as income, education, and internet access? All of this must be addressed when developing AI that is both spatially aware and linguistically inclusive. This chapter provides the conceptual basis for the empirical analysis presented in later chapters (Graham et al., 2015; Janowicz, 2023; Liu et al., 2025; Shi et al., 2025).

2.4 Sustainability and Environmental Cost of Inefficiency

The impact of AI on the environment is receiving more attention as researchers and policymakers recognize that AI computation has material consequences. LLMs require extensive computational resources for training and inference, which results in high energy and water consumption (Patterson et al., 2021; Strubell et al., 2019). Each token processed contributes to this incremental use of resources. Data centers are unevenly distributed, and electricity sources differ in carbon intensity (Shi et al., 2025). Water consumption is another major, yet less visible, cost metric, as data centers often rely on water for cooling, particularly in regions where evaporative systems are common. Li et al. (2023) provide estimates of the water footprint of AI systems, showing that interactions with large language models can be associated with measurable water consumption when indirect cooling and energy production are taken into account. In some cases, this consumption may reach the order of hundreds of milliliters per interaction. That means that linguistic inefficiency may be linked to differences in ecological impact, as users in low token efficiency regions indirectly require more computational resources for equivalent tasks (Shi et al., 2025).

Most hyperscale data centers are located in North America and Western Europe, where renewable energy sources and cooling technologies are well advanced and common. However, increasing demand from Asia, Africa, and Latin America is causing some of the ecological cost to shift to regions that rely on more intense carbon-based energy sources and that face water limitations. Shi et al. (2025) describe this transfer of environmental impact as a new form of environmental asymmetry in the digital age. Goodchild and Longley (2021) calls it the “geography of computation”, where global digital processes have physical and spatial consequences (Goodchild & Longley, 2021; Shi et al., 2025).

The energy consumption of LLMs is proportional to the number of computational operations executed per token. When a term is represented by tokens, both workload and

carbon emissions increase with increasing token count. Goldman et al. (2024) shows a correlation between token compression and energy efficiency: a more compact representation requires fewer model passes, resulting in reduced energy usage. Building on this insight, Wilhelm et al. (2025) introduces "energy per token" as a new evaluation metric for inference efficiency. The study compares energy usage across multiple models and reveals significant differences in "energy per token" output even for identical tasks. They argue that a sustainable AI evaluation should focus not only on total power usage, but also on the energy consumed for each token generated. This links with the concept of token efficiency, as it translates linguistic compactness into a measurable metric of environmental performance (Goldman et al., 2024; Wilhelm et al., 2025).

From a sustainability perspective, improving token efficiency goes hand in hand with the wider objective of reducing the carbon footprint of AI usage. Several strategies have been proposed. For example, compressing model vocabularies, adopting adaptive tokenization to minimize segmentation, or developing hardware-aware inference pipelines that can adjust batch sizes to token count. All of these approaches do target the early stages of the computational pipeline, where small gains in efficiency can have a big impact at inference time. Addressing inefficiencies in the tokenizer rather than in the transformer can result in environmental savings with minimal impact on quality (Ali et al., 2024; Rahman et al., 2024).

But optimizing tokenization for sustainability has its trade-offs. As extreme compression can bias token boundaries in favor of dominant languages, reducing representation for all other languages. Finding a way to achieve environmental efficiency without sacrificing linguistic inclusion remains a central challenge for the responsible design of AI (Shi et al., 2025).

Adopting a sustainable approach to token efficiency transforms language technology as part of the planetary infrastructure. Each additional token generates economic and ecological costs that increase by millions of interactions every day. Mapping these costs on a global scale, as this thesis proposes, provides a way to visualize the intersection of linguistic diversity, digital inequality, and environmental impact. This approach goes hand in hand with the principles of GeoAI. The proposed methods highlight regions where improving linguistic fairness can also improve environmental responsibility (Shi et al., 2025; Wilhelm et al., 2025).

2.5 Research Gap

Research in AI has produced detailed studies on topics such as tokenization, multilingual performance, ethics, and environmental sustainability. Linguistic works, such as Goldman et al. (2024), Rahman et al. (2024), and Petrov et al. (2023), describe how tokenization algorithms divide language and how this influences accuracy and cost. At the same time, sustainability research such as Shi et al. (2025) and Wilhelm et al. (2025) began to measure the resource requirements of training and inference, with a focus on energy input, carbon emissions, and water consumption. However, these fields remain disconnected. Tokenization is mostly examined as a technical or linguistic process, while sustainability and geography are viewed as external factors and not as an important component (Goodchild & Longley, 2021; Petrov et al., 2023; Rahman et al., 2024; Shi et al., 2025; Wilhelm et al., 2025).

Current studies do not explain how the linguistic structure and spatial context work together to shape the efficiency of language models. Although GeoAI research has shown that digital infrastructure and data access vary geographically (Goodchild & Longley, 2021; Janowicz, 2023; Liu et al., 2025), this has not yet been linked to tokenization. At the same time, environmental analyses do not consider the impact of language differences on resource use. As a result, a solid understanding of the relation between linguistic diversity, spatial inequality, and computational demand remains underexplored (Goodchild & Longley, 2021; Janowicz, 2023; Liu et al., 2025).

This thesis introduces token efficiency as a way to connect language structure, geography, and sustainability. Token efficiency is defined as the ratio between a minimal linguistic expression of meaning and the number of tokens required to represent it. The minimal expression refers to the shortest possible formulation of a place-related concept within a given language, while the token count reflects how this expression is segmented by a tokenizer. In this sense, token efficiency does not measure meaning itself, but how efficiently meaning can be encoded within a model. This allows tokenization to be interpreted as an indicator of relative efficiency across languages, which may have implications for computational cost. The thesis calculates this metric on a global scale using the H3 hexagonal grid and links token efficiency values to geographic regions and languages. By visualizing these values, it becomes possible to identify where higher or lower efficiency occurs and whether these patterns follow spatial structures (Birch et al., 2007; Brodsky, 2018; Janowicz, 2023; Wilhelm et al., 2025).

With this spatial perspective, the thesis connects linguistic variation and geographic patterns without assuming a direct causal relationship between them. Instead of treating tokenization purely as a technical step, it is considered in relation to where and how

language is used. This makes it possible to examine whether efficiency differences align with broader spatial patterns that have been described in GeoAI and digital inequality research (Goodchild & Longley, 2021; Liu et al., 2025; Shi et al., 2025).

In methodological terms, the thesis combines approaches from computational linguistics and spatial data science within a single analytical framework. Spatial methods are used to detect and describe geographic variation, while linguistic perspectives support the interpretation of structural differences in tokenization. References to sustainability are used to frame potential implications related to computational cost rather than to directly quantify environmental impact (Janowicz, 2023; Wang et al., 2025; Wilhelm et al., 2025).

By doing so, this thesis addresses the research gap by linking existing work on tokenization with geographic analysis, and by showing that efficiency differences in language representation can be observed as spatial patterns that emerge early in the processing pipeline of large language models.

3 Framework and Research Questions

This thesis builds on the theoretical foundations set out in the previous chapters. It connects linguistic, spatial, and environmental dimensions of LLMs through the concept of token efficiency and uses this concept as a measurable structure that can be tested empirically and visualized geographically.

3.1 Framework of the Thesis

Each token reflects how efficiently a model converts language into computation. When measured globally, token efficiency shows how these linguistic processes are unevenly distributed across space. Regions where models tokenize efficiently represent linguistic privilege, while low efficiency regions mark digital disadvantage. To study these patterns, this thesis adopts three perspectives:

1. Computational linguistics, following the footsteps of Sennrich et al. (2016), Rahman et al. (2024) and Goldman et al. (2024), explaining how sub-word segmentation influences model behavior.

2. GeoAI and spatial data science, providing the tools to map and analyze spatial structures in AI output. (Goodchild & Longley, 2021; Janowicz, 2023)

3. Sustainability research, building on the ideas of Shi et al. (2025) and Wilhelm et al. (2025) to link AI output to energy and water consumption.

Token efficiency serves as the common measure that links these domains.

3.2 Research Questions

Three main research questions are guiding this thesis. They build on the theoretical framework introduced in Chapter 3 and structure the empirical analysis.

RQ1: How much does token efficiency vary worldwide across different regions?

- What spatial patterns of efficiency can be observed on a global scale?
- Do regional clusters of high or low efficiency exist, and how are they distributed geographically?

RQ2: How do different tokenizers perform globally in terms of token efficiency?

- Are there differences in how consistent token efficiency is across regions between models?
- What spatial differences can be observed in their efficiency distributions?

RQ3: How are the observed patterns of token efficiency related to broader questions of digital inequality and sustainability?

- Do regions with differing levels of digital infrastructure show systematic differences in token efficiency?
- What potential implications do these patterns have for computational cost and resource use?

3.2.1 Scope of the Thesis

The scope of this thesis is global in coverage but selective in depth: it aims to reveal broad spatial and linguistic tendencies. The study explores how different LLMs tokenize geographical names around the world and how these differences form spatial patterns. The focus is on the distribution of efficiency and not on the individual language or the architecture of each model.

The analysis includes three tokenizers: the TikToken tokenizer accessed via the GPT-4o encoding, the tokenizer of Mistral-7B accessed through Hugging Face, and the tokenizer of DeepSeek-LLM-7B-Base, also accessed through Hugging Face. These models were selected because they each represent a different design philosophy and architecture.ChatGPT is the most widely used LLM in the world. It is optimized for English and developed in the USA. DeepSeek is a high-performance Chinese model, that was released to disrupt the market. It is completely free. While Mistral, on the other hand, was developed in Europe and stands for lightweight, open-access architecture. Together, they allow for meaningful cross-comparison of their spatial and linguistic behavior.

The study limits itself to token efficiency, which is defined as the ratio between linguistic form and token count. Other forms of model bias, like semantic accuracy and contextual fairness, are not analyzed. The metric does not measure the quality of a model, but efficiency in linguistic representation.

Spatially, this thesis operates at H3 resolution level 3, which balances computational feasibility with a meaningful geographic scale. Each cell aggregates a maximum of 300 geographic names. This leads to a capture of dominant linguistic structures while smoothing out local variation. The coverage includes all worldwide landmasses except Antarctica. Areas with insufficient OSM data coverage are left empty, and these gaps themselves reflect differences in digital visibility that are part of the broader pattern of digital inequality.

To contextualize the spatial patterns of token efficiency, four external data layers are included as comparison variables: population density, dominant writing system, carbon intensity of electricity, and internet penetration. These layers are examined descriptively and are intended to situate token efficiency within broader geographic, linguistic, environmental, and digital inequality structures. No causal relationships are assumed

The study period is static and not temporal. Data were collected at one point in time, creating a snapshot of the digital linguistic landscape of mid-2025. Changes over time, such as tokenizer updates or shifts in OSM coverage, are in fact outside the scope of this thesis. This work tries to be a baseline for future research rather than analyzing ever-evolving trends in AI.

Methodologically, this thesis introduces a workflow that can be reproduced, adapted to other models and data sources, or even spatial resolutions. It shows that token efficiency can be used as a metric to assess fairness in AI.

4 Methods

4.1 Model Selection

Table 4.1: List of LLMs

Model	Tokenizer Type	Developer	Open Source
GPT-4o	BPE (tiktoken)	OpenAI	✗
Claude 3.5 Sonnet	Unknown	Anthropic	✗
Gemini 1.5 Pro	SentencePiece	Google	✗
Mistral-7B	BPE variant	MistralAI	✓
LLaMA 3	SentencePiece (BPE)	Meta	✓
DeepSeek-LLM-7B	SentencePiece	DeepSeekAI	✓
Falcon 40B V2	BPE variant	UAE	✓
Yi 1.5 Large	BPE variant	01.AI	✓

The LLMs listed in Table 4.1 represent the major design approaches in the development of LLMs as of 2025. They differ in architecture, accessibility, and tokenizer. Open source models, such as Mistral, LLaMA, Falcon, and DeepSeek, offer access to their tokenizer code, leading to transparent replication of the tokenization processes. While commercial systems, such as ChatGPT, Gemini, and Claude expose to their tokenizers via APIs. Currently, ChatGPT is the only commercial model that allows its tokenizer to be used via the TikToken library.

From this group, three models were selected for this thesis: **GPT-4o, DeepSeek-LLM-7B and Mistral-7B**. Together they cover two of the most common tokenizer architectures used today: Byte-Pair Encoding and Unigram Language Modeling, and originate from the three major global centers of AI development: the United States, China, and the European Union.

GPT-4o is the commercial benchmark optimized for English and other major world languages (OpenAI, 2023). DeepSeek represents the class of open multilingual models designed for a broad linguistic coverage, including non-Latin scripts (DeepSeekAI, 2024). Mistral represents the generation of efficient and fully accessible European models that

have been trained on diverse multilingual datasets. (Jiang et al., 2024; Touvron et al., 2023)

4.2 Data Source

All geographic names were collected from OpenStreetMap (OSM) via the Overpass API. OSM was selected because it is a global, open-source, community-maintained dataset that provides human and physical geographic information with high global coverage, including data from almost all inhabited regions and languages worldwide. It also allows for easy reproducibility as it is open source and its categories, like 'place' or 'natural' correspond directly with our three semantic groups used in our sampling design. Finally, OSM captures linguistic diversity through locally contributed place names. As contributors often label features they live close to in their local language, which reflects the multilingual nature of the world better than official gazetteers.

Alternative datasets were tested, but they all proved to be unusable for performing a global analysis. GeoNames and HERE API, for example, provide detailed data in regions with strong digital infrastructure but have major gaps in Africa, Asia, and South America. Often, they were not suitable for finding more than a dozen names per cell in the local language.

4.3 Sampling Strategy

This thesis uses a stratified sampling strategy to ensure that each spatial cell represents a balanced and comparable selection of geographical names across predefined categories. The goal is to capture linguistic diversity while reducing bias introduced by differences in mapping density or geographic detail.

Each H3 cell is populated with place names sourced from OpenStreetMap according to three semantic categories:

- 100 Place names tagged with the category *"place"*. This includes the hierarchy of administrative boundaries. Examples of this are: Burgenland, Vienna, Donaustadt, Aspern, and so on.
- 110 Geographic features. This includes all natural elements such as rivers, mountains, lakes, or forests. The OSM tag for this is *"natural"* and *"waterway"*.

- 90 names of objects. This includes everything from buildings, points of interest, and infrastructure that has the OSM tag *"name"* but is without a specific geographic classification. Examples for this are: Stephansdom, Tangente, Universität Wien, or Staatsoper.

This distribution was chosen in order to achieve a maximum of 300 names per cell. It balances representational diversity and computational feasibility. A higher number would have increased the runtime without really further improving the linguistic diversity of each cell.

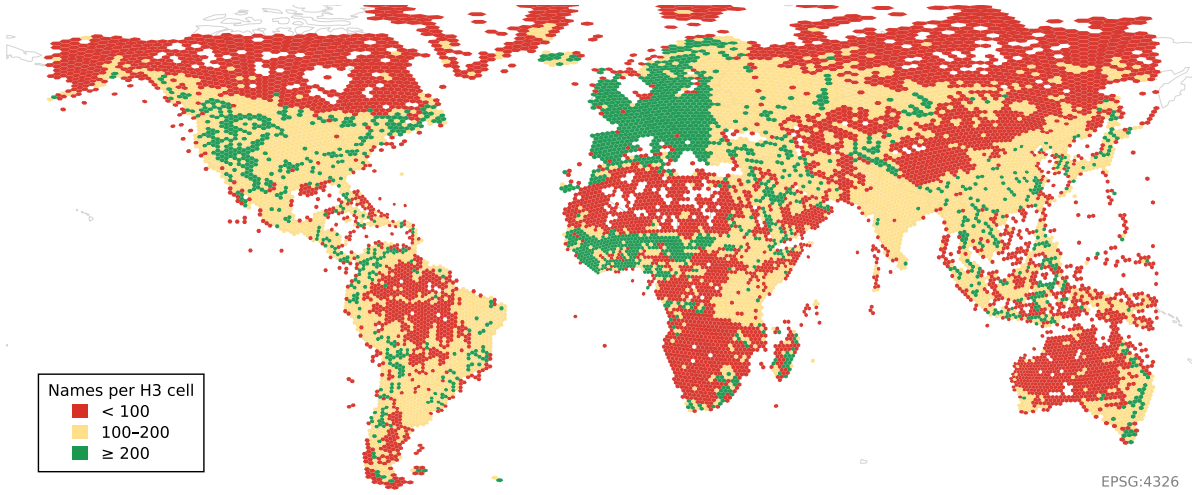


Figure 4.1: Global distribution of place name counts per H3 cell, grouped into three classes

Following initial data collection, only cells containing at least 100 valid names were included in the analysis. This threshold ensured that each included cell provided a sufficiently large sample to compute reliable token-efficiency averages. Cells with fewer than 100 entries were mainly located in sparsely mapped or low-population regions where OSM coverage is limited. Including these cells would have inflated the variance between cells. In total, out of 12580 cells, only 6911 cells met this criterion, and 5669 had to be removed. Figure 4.1 shows a map of name density indicating which cells were removed. Cells that are colored red contain fewer than 100 names, cells colored yellow between 100 and 200, and cells in green have more than 200 names. The resulting distribution reflects the strengths and gaps of global volunteered geographic information, especially the high level of completeness in urbanized and digitally connected regions, and the sparse coverage in parts of Africa, Central Asia, and Oceania.

This filtered dataset is the basis for all the following tokenization and analysis.

Table 4.2: Number of names and average number of names per H3 cell by category.

Category	Total Names	Avg. Names per Cell (incl. 0)	Avg. Names per Cell (excl. 0)
Natural	177335	14.10	22.83
Other	487893	38.75	52.44
Place	438626	34.87	46.19

4.4 Spatial Framework

The spatial framework for this thesis is based on the H3 hexagonal grid system developed by Uber. The H3 system divides the Earth´s surface into equal-area hexagonal cells with a hierarchical structure. Each hexagon is uniformly adjacent to its neighbors, which simplifies spatial analyses such as Moran´s I. This makes H3 perfect for a global comparative study that requires uniform spatial units and reproducible geometry. (Berrill et al., 2022; Brodsky, 2018; Mosa et al., 2025)

H3 is hierarchical, meaning that each increase in resolution subdivides every cell into seven smaller hexagons. With each level, the number of global cells grows exponentially. Table 4.3 summarizes the total number of cells for the first six resolutions and their geometric structure.

Table 4.3: H3 Resolution

Resolution	Total Number of cells	Avg area in km^2
0	122	4357449
1	842	609788
2	5870	86801
3	41162	12393
4	288122	1770
5	2016842	252

At resolution 3, the grid divides the world into 41,162 cells. This level provides a balance between spatial detail to detect regional variations in token efficiency and computational cost. Higher resolution would have expanded the dataset into unmanageable dimensions, as the JSON has nearly 1 GB at resolution 3.

4.4.1 Why H3

There are several discrete global grid systems, including Google´s S2. In S2, the Earth is projected onto the faces of a cube, with each face subdivided into smaller quadrilateral

cells. This design results in cells of unequal area and less uniform adjacency near cube edges. These irregularities could lead to biases when calculating spatial statistics or comparing regions. H3 mitigates some of these issues thanks to its equal hexagonal size and the natural neighbor structure, which reduces edge effects and distortion. (Birch et al., 2007; Goodchild et al., 2020)

H3 also has practical advantages for spatial analysis. While other grid systems such as S2 are also open source and supported by common programming libraries, H3 provides a hierarchical hexagonal grid that is particularly suited for spatial aggregation and neighborhood-based analysis. Hexagonal cells offer a more uniform representation of space, as they have equal distance to neighboring cells and reduce directional bias compared to square or rectangular grids. This is especially relevant for spatial statistics such as Moran's I and local clustering, where the definition of neighborhood relationships directly influences the results.

Berrill et al. (2022) shows that in large-scale environmental and spatial modeling, where uniform aggregation and neighborhood operations are essential, H3 performs effectively. Mosa et al. (2025) demonstrate that H3 outperforms several alternative discrete global grid system implementations in terms of scalability for global datasets and processing speed. In addition, the relatively uniform cell shapes of H3 improve comparability across regions in global analyses. This makes H3 well suited for the analysis of spatial patterns in token efficiency.

4.5 Token Efficiency

Token efficiency is measured as the relationship between the linguistic units of a geographic name and the number of tokens required by each model to represent it. Equation 4.1 shows the formulation used in this study. In this study, the minimum token count is defined as the number of words in a given place name, providing a consistent reference unit across the dataset. This yields a numerical value indicating how compactly each model represents text. A ratio close to 1 suggests high efficiency, while lower values point to fragmentation. The measure reflects how linguistic units are segmented into tokens by each model. It also enables comparisons between languages and tokenizer types, while still reflecting differences in linguistic structure.

$$\text{Token Ratio} = \frac{\text{Minimum Token Count}}{\text{Actual Token Count}} \tag{4.1}$$

For each name extracted from OSM, token efficiency was calculated separately using the tokenizers GPT-4o, DeepSeek-LLM-7B, and Mistral-7B. Tokenization was performed using the official Python libraries for each model. As each tokenizer uses different segmentation logic, a direct comparison can be made between those approaches under identical conditions. (Rahman et al., 2024; Sennrich et al., 2016)

The resulting token counts were aggregated at the H3 cell resolution 3. For each cell, the mean and standard deviation of the token ratio were calculated for all names in the cell. This resulted in a dataset of average token efficiency for each model and cell, which was used as input for all future statistics and maps.

An example of this method can be demonstrated as follows: The German name "Wien" is encoded by GPT-4o tokenizer as one token, producing a token ratio of exactly 1. Meanwhile, the small Austrian village named "Pfaffenschlag bei Waidhofen an der Thaya" is divided into 12 tokens. This results in a ratio of 0.5, meaning that 12 tokens are required to represent a name with 6 words. A Chinese place name such as "北京" is treated as a single word and is split into 2 tokens, producing a 0.5 ratio.

5 Results

5.1 Basic Statistics

Table 5.1 summarizes the global statistics for all cells. GPT-4o has a mean token ratio of 0.412 and a standard deviation of 0.153; DeepSeek-LLM-7B has an average of 0.333 and a standard deviation of 0.106; and Mistral-7B has the lowest mean of 0.310 and a standard deviation of 0.102. The minimum and maximum values follow the same relative order: GPT-4o achieves the broadest range, while Mistral-7Bs values are the most compressed.

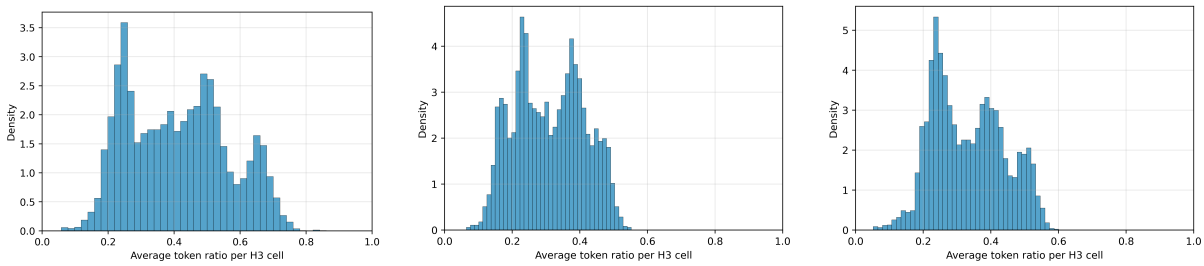
Table 5.1: Descriptive statistics of average token ratios per H3 cell for the three evaluated models.

Model	**Mean**	**Std. Dev.**	**Min**	**Max**	**Cells (n)**
GPT-4o	0.412	0.153	0.058	0.861	6911
DeepSeek-LLM-7B	0.333	0.106	0.052	0.601	6911
Mistral-7B	0.310	0.102	0.064	0.552	6911

This first comparison highlights two key points. Firstly, all three tokenizers share a structural hierarchy of efficiency, with GPT-4o showing the highest token efficiency, followed by DeepSeek-LLM-7B and Mistral-7B. This suggests that the differences encountered through this thesis are systematic rather than random. Secondly, GPT-4o´s high standard deviation value shows that it performs very efficiently in some regions but moderately elsewhere; Mistral-7B, meanwhile, remains more consistent but is less efficient overall for tokenizing place names. DeepSeek-LLM-7B falls in between, showing solid compression but a narrower range.

The three histograms in Figure 5.1a, Figure 5.1c and Figure 5.1b visualize the global distribution of token ratios for the three evaluated models. Each histogram uses 40 equally sized bins.

Figure 5.1a shows that the token ratio for GPT-4o ranges between 0.058 and 0.861, with two visible peaks at roughly 0.25 and 0.5. The density rises sharply around 0.2 and flattens above 0.55. A local peak near 0.65 indicates that a smaller number of cells reach relatively



(a) GPT-4o (b) Mistral-7B (c) DeepSeek-LLM-7B

Figure 5.1: Distribution of average token ratios per H3 cell for the three evaluated LLMs. Each histogram shows the density of the average token ratio aggregated per H3 cell ($n \geq 100$).

high efficiency values. This shape suggests that GPT-4o exhibits considerable variability. Its tokenizer achieves very compact representations in some regions, while in others it is noticeably less efficient. The widespread and multiple peaks indicate heterogeneous global performance.

Figure 5.1c shows a smaller range, from 0.05 to 0.6, with most values concentrated between 0.2 and 0.45. The histogram is more compact and slightly left-skewed, indicating that many cells have lower efficiency values. The density decreases beyond 0.5 and does not extend as far as the GPT-4o distribution. This suggests that DeepSeek-LLM-7B produces fewer very high efficiency values and instead concentrates more observations in the lower and middle ranges. While the model appears more stable, it rarely reaches the same level of compression as GPT-4o.

Figure 5.1b shows a range of 0.065 to 0.55, with two major peaks around 0.225 and 0.4. The overall shape of the histogram is smoother and more symmetric than that of the other two models. Values above 0.5 are rare, and no extreme outliers are present. This indicates that Mistral-7B produces values within a narrower range, with most observations concentrated in the lower to mid efficiency levels. This suggests a more compact distribution compared to the other models.

To further compare the distributions, Figure 5.2 presents a violin and boxplot representation of the same data. While the histograms highlight the frequency distribution, the violin plots emphasize the overall shape and density of the token-ratio distributions across models. The width of each violin represents the concentration of observations at a given value, while the embedded boxplots indicate the median and interquartile range.

The violin diagram confirms the patterns observed in the histograms. GPT-4o exhibits the widest distribution, with a broader spread and higher maximum values, suggesting that it occasionally achieves very compact tokenizations of place names. DeepSeek-LLM-7B shows a narrower distribution centered on lower-to-mid-range values, suggesting a

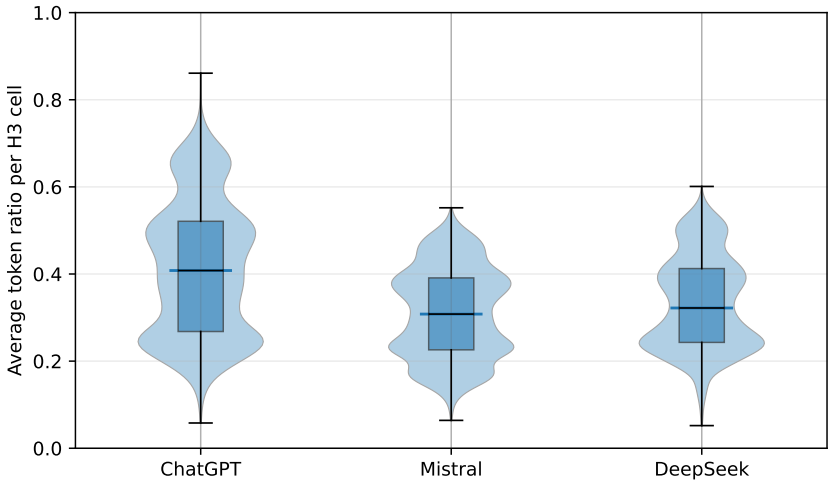
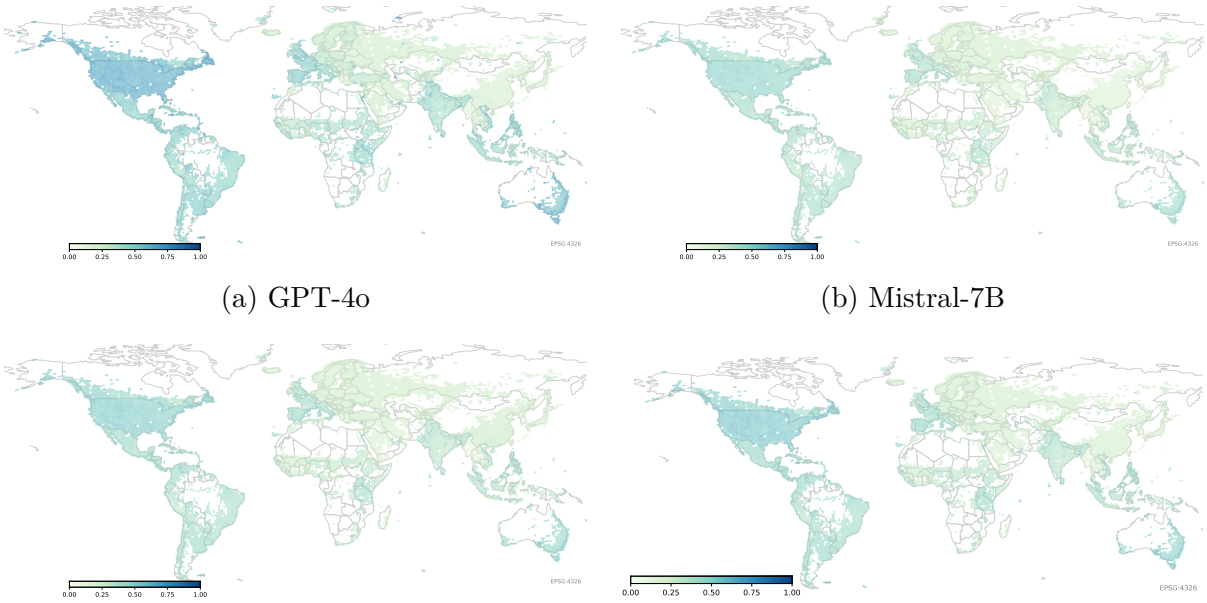


Figure 5.2: Distribution of average token ratios per H3 cell for GPT-4o, Mistral-7B, and DeepSeek-LLM-7B (values clipped to $[0, 1]$, outliers hidden). Violin plots show distribution shape; boxplots show median and interquartile range.

more compact but generally lower efficiency. Mistral-7B presents the most compact and symmetrical distribution, indicating the most consistent behavior across cells, although it did not reach the highest efficiency levels observed for GPT-4o.



(a) GPT-4o

(b) Mistral-7B

(c) DeepSeek-LLM-7B

(d) Mean of three models

Figure 5.3: Average token ratio per H3 cell for GPT-4o, Mistral-7B, and DeepSeek-LLM-7B, and their mean across all three models (EPSG:4326; n 100).

However, what truly matters is the geographic characterization. Figure 5.3 shows the global distribution of token efficiency, mapping the mean token ratio per H3 cell per tokenizer and aggregated for all three tokenizers. Dark shades of blue indicate higher efficiency, where names are encoded with relatively few tokens, while lighter shades mark

low(er) efficiency.

The pattern is clearly spatial and structured. Similar values are clustered into large regions that form clear macro-zones rather than appearing scattered. Regions with high efficiency are concentrated in all English-speaking parts of the world, like North America, the British Isles, and Oceania. All Regions, where Romance Languages are spoken, such as Latin America, Western- and parts of Southern Europe or Quebec, also have relatively high efficiency.

Moderate efficiency regions are located in some parts of Eastern Asia, the Indian Subcontinent, and around Lake Victoria in Africa. These zones appear in medium shades of blue and are often less homogeneous than high-efficiency regions.

Large parts of Sub-Saharan Africa, Central Asia, and Eastern Europe are dominated by low-efficiency areas. Their values fall well below the global mean. The map shows that low efficiency is not spatially isolated, with entire areas extending thousands of kilometers sharing similar values. The smooth transitions between efficient and inefficient areas demonstrate that tokenization follows a (geo)spatial logic, where neighboring areas tend to behave similarly.

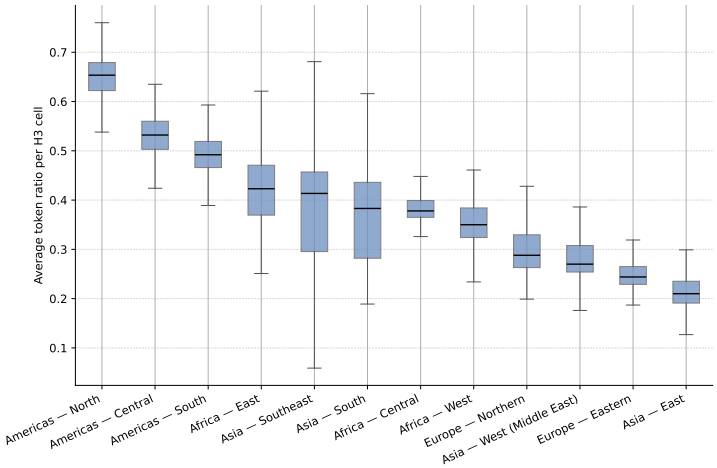
Figure 5.4a, Figure 5.4b, and Figure 5.4c show the distribution of token efficiency across macro regions. Each box plots the spread of the average token ratio per macro region, while the orange line marks the median. While the models have different overall efficiency levels, their regional hierarchies are almost identical.

North America has the highest median efficiency in all three models. The boxes are narrow, with the upper quartiles reaching the top of each plot, indicating stable and consistent performance. Oceania follows closely behind and displays similar levels of consistency. These two macro regions form a clear global peak of token efficiency.

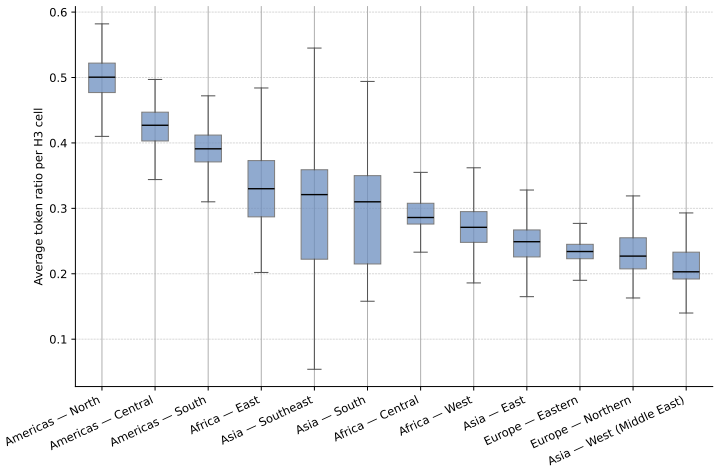
The Middle East ranks near the bottom for all three models. The medians are low, and the quartile ranges are tight, showing uniform weakness across tokenizers. A similar pattern is recognizable in Eastern Europe, where the boxes are small and positioned at low values.

East Asia stands out for its low values and clear differences between models. GPT-4o achieves its lowest global median in this region, performing notably worse than the other two tokenizers. DeepSeek-LLM-7B performs better than both GPT-4o and Mistral-7B in East Asia, but loses efficiency in Europe, with the lowest values in northern and eastern Europe.

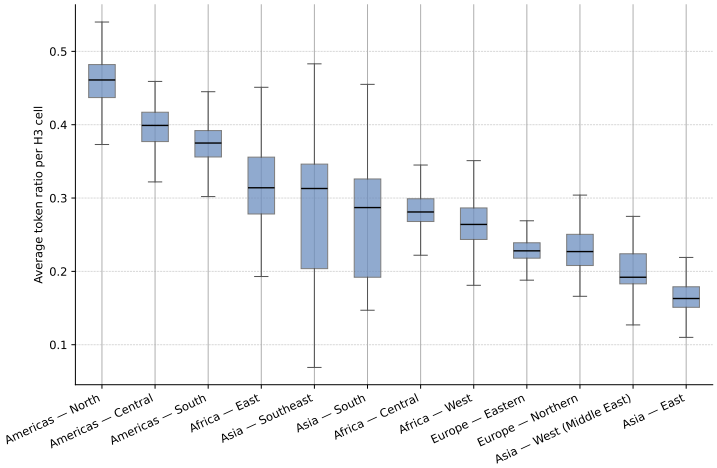
Out of all regions, South East Asia shows the widest spread. Each model has strong outliers, both high and low, suggesting large variations between cells in that region. This



(a) GPT-4o



(b) DeepSeek-LLM-7B



(c) Mistral-7B

Figure 5.4: Distribution of average token ratios per H3 cell across macro regions for the three evaluated tokenizers.

means that token efficiency is highly uneven, with some cells performing close to the global average and others dropping to the lowest observed values. Unlike the stable clusters observed in North America or East Asia, South East Asia does not form a coherent zone of similar values. South Asia, which is the Indian Subcontinent, has the widest interquartile range across all three models, suggesting a comparable mixture of higher and lower efficiency areas.

Table 5.2: Top ten countries with the lowest overall token ratio across all evaluated models.

Country	**Mean Token Ratio**	**Cells (n)**	**Std. Across Models**	**Range Across Models**
South Korea	0.157	13	0.044	0.089
Japan	0.159	49	0.020	0.037
Greece	0.162	14	0.016	0.032
Myanmar	0.166	56	0.027	0.047
North Korea	0.185	18	0.046	0.092
Georgia	0.197	8	0.003	0.007
Thailand	0.203	45	0.047	0.085
Iran	0.208	120	0.043	0.083
Lithuania	0.215	8	0.018	0.032
Iraq	0.216	29	0.039	0.072

To provide a better picture of spatial variation, the token efficiency results were aggregated at the country level. The 10 best-performing countries reveal a clear dominance of English and Spanish-speaking countries. The United States, Canada, and Australia have the highest mean values across all models, followed by Latin American countries like Nicaragua, Costa Rica, and Mexico. All of these countries are in the Top 10 in all three models. After that, other Spanish-speaking countries like Venezuela, Cuba, and Spain follow in different rankings. Their average token ratios range from USA's 0.66 for GPT-4o to Honduras and its value 0.385 for Mistral-7B. All clearly above the global mean for each tokenizer.

At the lower end of the global distribution, as Table 5.2 shows, countries like South Korea, Japan, Greece, Myanmar, Georgia, Thailand, and Iran can be found for every model. These countries have mean ratios below 0.22, meaning their place names require more than twice as many tokens as those of the highest-ranked countries. There is little variation between models in these regions, showing that inefficiency is consistent across model types.

Further analysis of the lowest-ranked countries shows several patterns. South Korea and Japan have low means with little to no variation, indicating uniform inefficiency. Greece,

Myanmar, and North Korea can also be placed into that group, but they show slightly higher spreads, which might come from edge effects. The ten worst-performing cells of the whole dataset are, for example, all located in the middle of Myanmar, with only names from their local language. Table 5.3

Table 5.3: Top 10 H3 Cells with Lowest GPT-4o Token Ratio

Cell	Lat	Lon	#Names	Names (Preview)
8364aafffff	16.408490	96.890864	144	မင်္လာပါ□ ဘတ်□ ဆောင်မကာန်□ □□□
83648dfffff	16.934236	94.843081	104	စောဂီနီ □□□□ ရေအဝတ် □□□□ ဧရာ □□□□ □□□
836481fffff	17.530042	97.215120	162	သဘော်ာ □□□□ ကားဂ □□□□ ပုံတော် □□□□ □□□
833cd8fffff	20.780975	93.728459	190	အောင်ဒေး □□□□ မြို့မ □□□□ စာသင်ကျောင်း □□□□ □□□
83648effffff	18.056310	95.160537	110	ကံသီ □□□□ ရေကျေး □□□□ တောင်ကြီး □□□□ □□□
83648cffffff	17.235800	96.027932	110	တောင်ကြီး □□□□ အင်းချောင် □□□□ တောင်သာ □□□□ □□□
833cdfffffff	19.673641	93.413838	114	မောင်လှ □□□□ အထက်တောင် □□□□ ဗဟန်း □□□□ □□□
836414fffff	11.629749	98.796403	226	သဲခေါင် □□□□ ရေကြောင်းလမ်း □□□□ အရပ် □□□□ □□□
832c2bfffff	42.246901	45.718509	239	Կորպ □□□□ Հալաբ □□□□ Արթուր □□□□ Տերտեր □□□□ □□□
833ccafffff	22.179760	95.248063	190	နောင် □□□□□ မြို့မ □□□□ ရေတောင်ကြီး □□□□ □□□

5.2 Cross Model Correlation

To measure model similarity in behavior across geographic space, the mean token ratio per H3 cell was compared across all three models. The resulting correlation matrix shows values above 0.95 for every model pair. The strongest correlation is between GPT-4o and Mistral-7B at 0,974, followed closely by DeepSeek-LLM-7B and Mistral-7B at 0,973 and GPT-4o and DeepSeek-LLM-7B at 0.964. These extremely high and close values show that the three tokenizer produce almost identical spatial efficiency structures.

This means that, although the models have differing average performance levels, they rank the world´s regions in almost the same way. Areas that perform well in one model are highly likely to also perform well in others, and areas that perform poorly tend to remain so regardless of the tokenizer. This common spatial structure shows that token efficiency is driven by systematic characteristics of how text is segmented and represented across languages and not by model-specific behavior.

Table 5.4: Pairwise correlation of mean token ratios per H3 cell across the evaluated models.

Model	GPT-4o	DeepSeek-LLM-7B	Mistral-7B
GPT-4o	1.000	0.964	0.974
DeepSeek-LLM-7B	0.964	1.000	0.973
Mistral-7B	0.974	0.973	1.000

The high correlation also explains why the spatial patterns observed in the global maps and macro-regional box plots remain consistent across models. Even when the absolute efficiency values differ, the relative positions of regions do not change. That means that differences between models appear mainly as shifts in level rather than changes in spatial structure. Regions that form high-efficiency clusters in one model also form them in the others.

5.3 Spatial Autocorrelation

In order to test whether the observed spatial patterns of token efficiency are statistically significant, a global Moran´s I was calculated for each tokenizer, based on the mean token ratio per H3 cell. This was done to test whether similar values occur close together or are randomly distributed across space.

As seen in Table 5.5 all three models show very high statistically significant Morans I values: GPT-4o has a Moran´s I value of 0.92, DeepSeek-LLM-7B of 0.917 and Mistral-7B of 0.924. The associated p-value for all three models is 0.001, showing that the probability of observing this degree of spatial clustering under a random spatial distribution is low.

These results confirm that token efficiency is strongly spatially clustered. Cells with high token ratios tend to be located near other high-value cells, while cells with low efficiency tend to be surrounded by similarly low-value cells. That means that the spatial structure visible in the global maps is not only visually existent but also a statistically robust pattern.

The similarity of Moran's I values across all three models indicates that the strength of spatial clustering is nearly identical between the tokenizers. Even though the models have different average efficiency levels, they display the same degree of spatial dependence, suggesting that the spatial organization of token efficiency remains consistent across tokenizers.

Table 5.5: Global Moran's I statistics for average token ratios per H3 cell across the evaluated models.

Model	Moran's I	p-value
GPT-4o	0.920	0.001
DeepSeek-LLM-7B	0.917	0.001
Mistral-7B	0.924	0.001

5.4 Local Spatial Autocorrelation

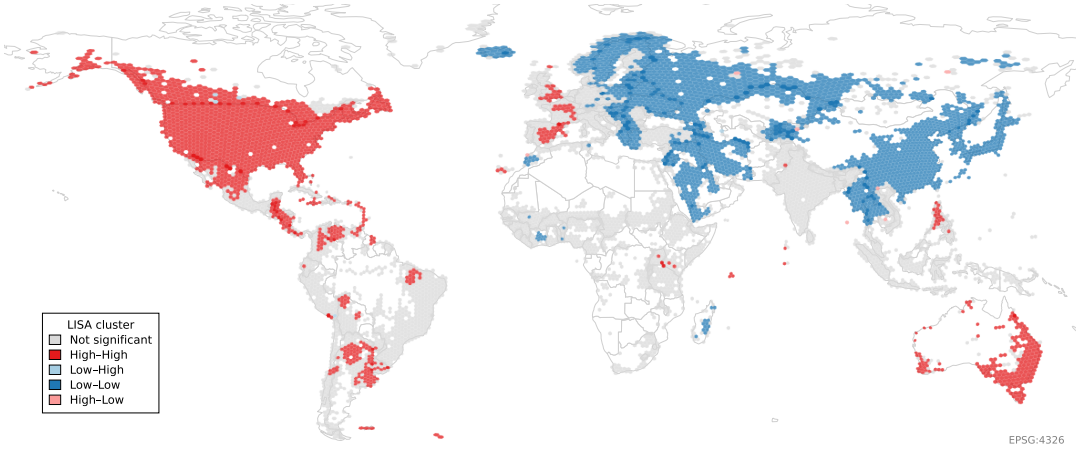
The LISA maps reveal clear regional cluster structures across all three models. High-high clusters primarily occur in North America and parts of Oceania, indicating areas where relatively high token efficiency values occur alongside similarly high values. These regions form continuous zones rather than isolated cells, reflecting the strong global spatial autocorrelation already detected through Moran's I.

In contrast, low–low clusters are concentrated in large parts of Asia and in several regions of Eastern Europe. In these areas, cells with lower token efficiency values are surrounded by neighbors with similarly low values. The extent of these clusters suggests that reduced token efficiency is not limited to isolated locations but rather forms larger regional patterns.

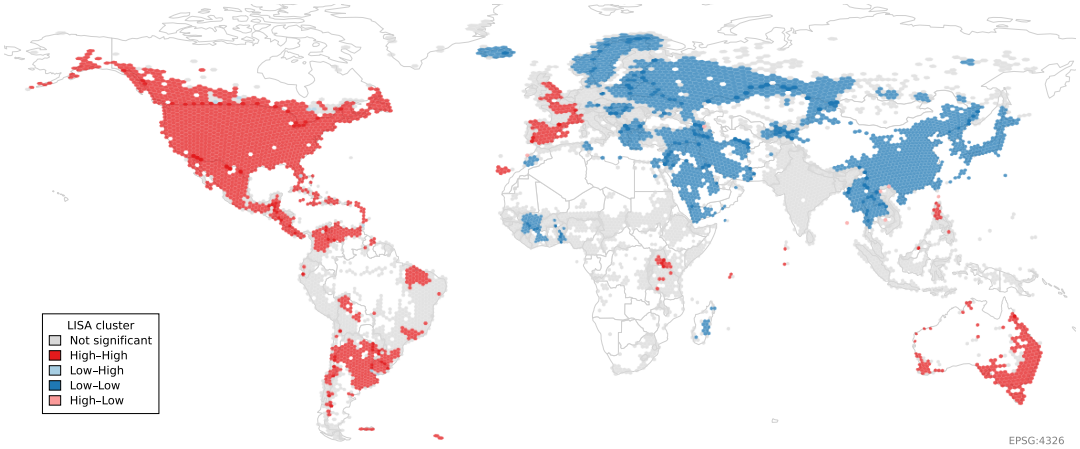
Some regions, particularly Southeast and South Asia, exhibit a more fragmented structure. Rather than forming large continuous clusters, these areas contain a mixture of significant clusters and spatial outliers. This indicates a stronger local variation in token efficiency, with neighboring cells differing more from each other than in the more stable cluster regions.

A large proportion of cells are classified as not statistically significant. This indicates that, although token efficiency is spatially autocorrelated at a global level, local clustering is concentrated in distinct regions, while large areas show no statistically significant local structure. As such, the spatial structure emerges through a small number of larger clusters rather than a uniform global pattern.

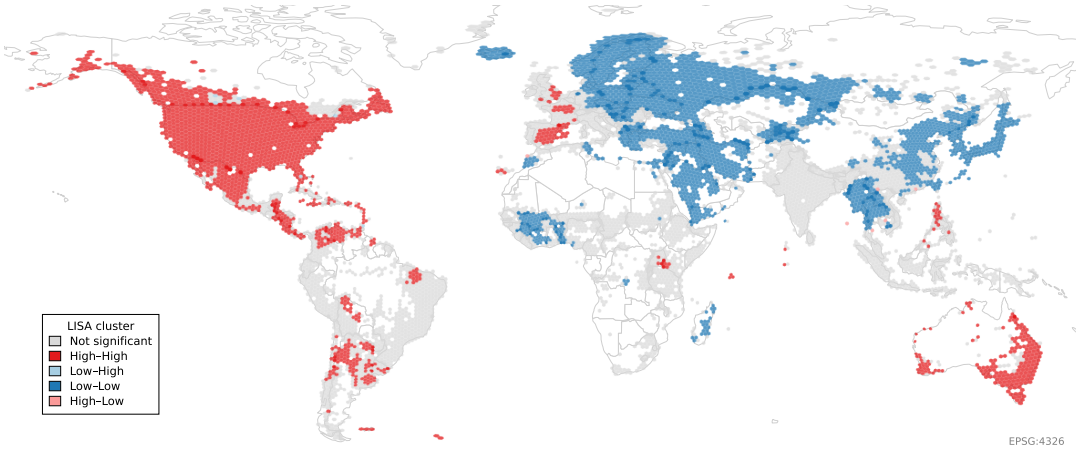
Overall, the cluster locations remain consistent across the three tokenizer models. Regions that form high–high or low–low clusters in one model tend to appear in the same category in the others. This reinforces previous findings that differences between tokenizers primarily impact the absolute efficiency level, while the spatial structure of token efficiency remains largely unchanged.



(a) GPT-4o



(b) Mistral-7B



(c) DeepSeek-LLM-7B

Figure 5.5: Local Moran's I (LISA) cluster maps of average token ratios per H3 cell (EPSG:4326; $\alpha = 0.05$; $n \geq 100$).

5.5 Rank Stability Across Models

To further examine whether the observed similarities between models exist beyond cell-level values, Spearman´s rank correlation was calculated using the country-level mean token ratios. Spearman focuses on the relative ordering of countries rather than their absolute efficiency values.

The rank correlations are very high for all models. As seen in Table 5.6 GPT-4o and DeepSeek-LLM-7B reach a Spearman correlation of 0.979; GPT-4o and Mistral-7B also reach 0.979; and DeepSeek-LLM-7B and Mistral-7B reach 0.988. These values clearly show that countries are ranked almost identically across all three tokenizers.

Table 5.6: Pairwise Spearman correlation of mean token ratios per H3 cell across the evaluated models.

Model	**GPT-4o**	**DeepSeek-LLM-7B**	**Mistral-7B**
GPT-4o	**1.000**	0.979	0.979
DeepSeek-LLM-7B	0.979	**1.000**	0.988
Mistral-7B	0.979	0.988	**1.000**

The country rankings presented in Table 5.7 and Table 5.8 further support the patterns observed in the previous analyzes. At the upper end of the distribution, the same group of countries consistently appears among the highest performing cases across all three models. In particular, the United States, Canada, and Australia are ranked among the top countries for GPT-4o, DeepSeek-LLM-7B, and Mistral-7B. Although the absolute token ratios differ between the models, the relative position of these countries remains very similar across all rankings.

A comparable pattern can also be observed among the countries with the lowest efficiency values. Several countries appear repeatedly in the bottom part of the rankings across the different tokenizers. Japan, South Korea, Greece, Myanmar, and Taiwan are consistently located among the least efficient countries in all three models. While the exact token ratios vary slightly, their relative positions within the rankings remain largely unchanged.

These observations indicate that differences between tokenizers mainly influence the absolute efficiency values rather than the relative ordering of countries. Therefore, the national distribution of token efficiency appears to be largely stable across models. Countries that achieve comparatively high efficiency with one tokenizer tend to perform similarly to the others, while countries with lower efficiency remain near the bottom of the rankings regardless of the tokenizer used.

Table 5.7: Top ten countries with the highest average token ratios per H3 cell for the three evaluated models.

Country	**GPT-4o**	**DeepSeek-LLM-7B**	**Mistral-7B**
United States of America	0.661	0.507	0.467
Canada	0.603	0.460	0.429
Australia	0.595	0.461	0.426
Guyana	0.587	0.438	0.410
Nicaragua	0.567	0.436	0.410
Fiji	0.555	0.412	0.395
Costa Rica	0.541	0.428	0.399
Guatemala	0.529	–	–
Mexico	0.528	0.425	0.397
Venezuela	0.523	0.421	0.398
Spain	–	0.421	0.401

Table 5.8: Ten countries with the lowest average token ratios per H3 cell for the three evaluated models.

Country	**GPT-4o**	**DeepSeek-LLM-7B**	**Mistral-7B**
Japan	0.174	0.167	0.137
Taiwan	0.186	0.186	0.143
Greece	0.176	0.164	0.144
Myanmar	0.197	0.152	0.149
South Korea	0.203	0.115	0.154
North Korea	–	0.144	0.177
Thailand	–	0.172	0.180
Georgia	0.201	0.194	–
Iran	–	–	0.172
People's Republic of China	0.224	–	0.172
Israel	0.224	0.194	0.184
Belarus	0.229	–	–
North Macedonia	0.231	–	–
Hungary	–	0.196	–

5.6 Data Layer Comparison

5.6.1 Population Density

Population density was selected as an external data layer to examine whether the spatial distribution of token efficiency is associated with the distribution of human population. The preceding sections show that low efficiency regions are concentrated in parts of East Asia, Southeast Asia, and the Middle East. These regions also contain some of the most densely populated areas in the world. Whether this overlap is systematic cannot be determined from the spatial maps alone. A quantitative comparison is needed to assess whether population density and token efficiency co-vary across cells. This comparison is also relevant from a digital inequality perspective, as a tokenization disadvantage concentrated in densely populated regions affects a larger number of users than one concentrated in sparsely populated areas.

Population data were obtained from the GHS Population Grid 2025, produced by the European Commission Joint Research Center (JRC), at a spatial resolution of 1km (European Commission, Joint Research Centre, 2023). For each H3 cell, zonal statistics were computed by aggregating all raster pixels falling within the cell polygon, producing a mean population density in people per km^2 and a total population sum per cell.

Table 5.9: Descriptive statistics of population density and population sum per H3 cell derived from the GHS Population Grid (2025). Based on 6,909 unique H3 cells with valid population estimates.

Statistic	**Population Mean (per km²)**	**Population Sum**
Mean	113.33	1,115,100
Std	260.61	2,521,214
Min	0.00	0
25%	5.69	53,169
50%	29.96	292,280
75%	106.46	1,031,956
Max	7,503.42	42,684,470

Population values vary strongly across H3 cells, as shown in Table 5.9. The median population sum per cell is 296,528, while values range from 0 to more than 42 million. A similar pattern is visible for mean population density, with a median of 29.75 people per km^2 and a maximum of 7,503 people per km^2. Both distributions are strongly right-skewed, indicating that high population values are concentrated in a small number of cells.

The association between token efficiency and population was examined using Spearman rank correlation between the population sum per cell and the mean token ratio for each model. As shown in Table 5.10, all three models show a weak but statistically highly significant negative correlation. The coefficients range from -0.186 for DeepSeek-LLM-7B to -0.247 for Mistral-7B, with p-values below 0.001 in all cases, indicating that cells with higher population values are associated with slightly lower token efficiency values. The correlation values are consistent in direction across all three models but remain weak, suggesting that population sum is only one of several factors associated with the spatial distribution of token efficiency.

Table 5.10: Spearman rank correlation of mean token ratios per H3 cell against population mean density and population sum derived from the GHS Population Grid (2025). Based on 6,909 unique H3 cells.

Model	**r (Pop. Mean)**	**p**	**r (Pop. Sum)**	**p**
GPT-4o	-0.209	< 0.001	-0.222	< 0.001
DeepSeek-LLM-7B	-0.174	< 0.001	-0.186	< 0.001
Mistral-7B	-0.235	< 0.001	-0.247	< 0.001

5.6.2 Writing System

Writing system was selected as a second comparison layer to examine whether the spatial patterns of token efficiency are associated with the dominant script of place names within each H3 cell. This comparison is motivated by the observation that subword tokenization algorithms are trained predominantly on text from Latin script languages, and that morphological and orthographic differences between writing systems may influence how efficiently text is segmented into tokens. (Petrov et al., 2023; Rahman et al., 2024)

The dominant script of each H3 cell was detected directly from the place name strings stored in the dataset, using Unicode character classification. For each character in each place name, the Unicode block was identified and assigned to one of the following categories: Latin, Cyrillic, ChineseJapanese (CJ), Arabic, Hangul, Thai, Myanmar, Greek, Hebrew, Georgian, Armenian, or Devanagari. The dominant script of each name was determined by the most frequent script among its characters. The dominant script of each cell was then determined by majority vote across all place names in that cell. Within the Latin category, two subtypes were distinguished based on the presence of distinctive diacritic characters: Latin Vietnamese, identified by the presence of Vietnamese specific diacritical marks, and Latin Turkish, identified by the presence of Turkish specific characters such as ğ, ş, and ı. All remaining Latin script cells were classified as Latin Basic.

Script groups with fewer than 20 cells were excluded from the analysis. The final dataset used for this analysis consists of 6,884 unique cells across ten script groups.

Latin Basic is the largest group with 4,892 cells, reflecting the global coverage of Latin script across the Americas, Europe, Africa, and parts of Asia. Cyrillic covers 748 cells concentrated in Russia and Central Asia. CJ covers 606 cells in East Asia. Arabic covers 399 cells across the Middle East and North Africa. The remaining groups are smaller, with Myanmar, Hangul, Thai, Greek, Latin Turkish, and Latin Vietnamese each covering between 24 and 65 cells.

The distribution of token efficiency across script groups is shown in Figure 5.6. Latin Basic shows the highest median token ratio across all three models and the widest interquartile range, reflecting the linguistic diversity of cells classified under this group. Latin Vietnamese follows with a similarly wide distribution, though its median is noticeably lower than Latin Basic in all three models. Latin Turkish shows a narrower and more compact distribution, positioned below both Latin groups.

Arabic and Cyrillic occupy a middle range across all three models, with relatively compact distributions and similar median values. Thai, Hangul, CJ, Greek, and Myanmar are all concentrated in the lower part of the distribution, with narrow interquartile ranges indicating limited variation within these groups. Myanmar shows the lowest median values across all three models and the most compact distribution, with values concentrated near the bottom of the observed range. One pattern that differs across models is visible in the CJ group. For GPT-4o and Mistral-7B, CJ shows a median and interquartile range similar to Hangul and Thai, placing it among the lower efficiency groups. For DeepSeek-LLM-7B, the CJ box is positioned noticeably higher, with a median closer to Arabic and Cyrillic. This difference is consistent with the cross-model patterns identified in earlier sections, where DeepSeek-LLM-7B showed slightly distinct behavior in East Asian regions.

The script group analysis indicates that the spatial distribution of token efficiency is closely associated with the dominant writing system per cell. The regions identified as low efficiency in earlier sections, particularly East Asia, Southeast Asia, the Middle East, and parts of Eastern Europe, correspond to cells dominated by CJ, Thai, Hangul, Arabic, and Cyrillic scripts respectively. The regions identified as high efficiency, primarily the Americas, Western Europe, and Oceania, correspond predominantly to Latin Basic cells. This suggests that the spatial patterns of token efficiency observed at the global level are to a substantial degree structured by the geographic distribution of writing systems rather than by model specific design choices.

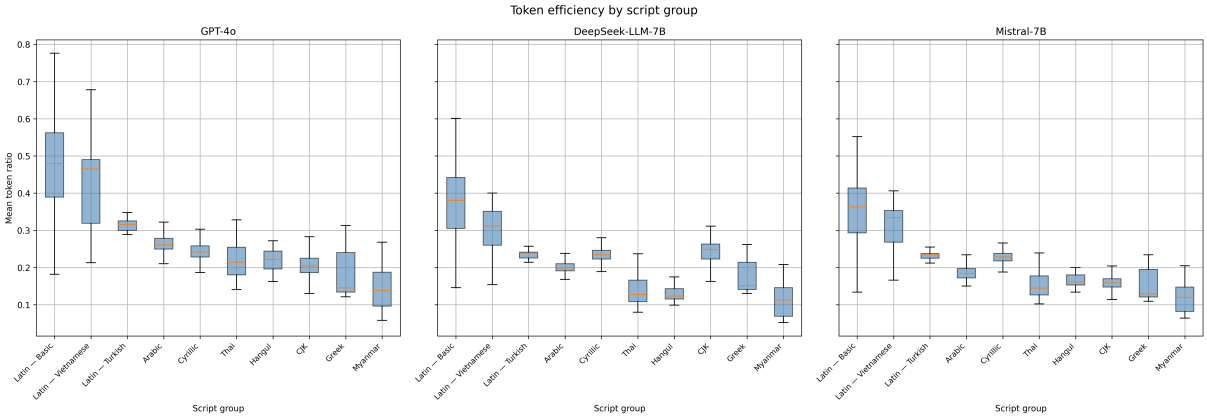


Figure 5.6: Distribution of mean token ratios per H3 cell across dominant script groups for GPT-4o, DeepSeek-LLM-7B, and Mistral-7B. Box plots show the interquartile range and median. Outliers are not shown. Script groups with fewer than 20 cells are excluded.

5.6.3 Carbon Intensity of Electricity

Carbon intensity of electricity was selected as a third comparison layer to examine whether the spatial distribution of token efficiency is associated with the environmental cost of computation. Regions with low token efficiency require more tokens to process the same content. If these regions also rely on carbon intensive electricity grids, the computational cost per unit of text is higher in both token count and associated emissions. This combination motivates a cell-level index that captures both dimensions together. Carbon intensity data were obtained from Our World in Data, sourced from Ember and the International Energy Agency, and represent national electricity grid emissions in grams of CO2 per kilowatt hour for the year 2022 (Ember & International Energy Agency, 2022).

$$\text{Environmental Cost Index} = \frac{1}{\bar{r}} \times C \tag{5.1}$$

where $\bar{r}$ is the mean token ratio of the H3 cell and C is the national carbon intensity of electricity in gCO_2 per kWh. A lower value of $\bar{r}$ indicates lower token efficiency, resulting in a higher index value.

As seen in Equation 5.1 the index is calculated by multiplying the inverse of the mean token ratio by the national carbon intensity value in grams of CO2 per kilowatt hour. A higher inverse token ratio indicates lower token efficiency, meaning more tokens are needed per unit of text. Multiplying this by carbon intensity produces a relative estimate of how much carbon-intensive computation is associated with processing place names in a given cell. The index is relative and does not represent an absolute measure of energy

consumption or emissions. It is intended as an exploratory indicator that combines two independently measured variables into a single comparable value across cells.

The association between token efficiency and carbon intensity was further examined using Spearman rank correlation. All three models show a moderate negative relationship, with coefficients of -0.367 for GPT-4o, -0.332 for DeepSeek-LLM-7B, and -0.391 for Mistral-7B, all statistically highly significant ($p < 0.001$). This indicates that cells with lower token efficiency values tend to be located in countries with higher carbon intensity electricity grids. The correlation values are consistent in direction across all three models.

Table 5.11: Descriptive statistics of the relative environmental cost index per H3 cell for all three models. The index is computed as the product of token inefficiency and national carbon intensity of electricity (gCO2 per kWh, Our World in Data 2022). Based on 6,290 matched cells.

Statistic	**GPT-4o**	**DeepSeek-LLM-7B**	**Mistral-7B**
Mean	1,292.98	1,504.04	1,665.40
Std	992.84	1,094.23	1,232.24
Min	0.00	0.00	0.00
25%	573.97	745.01	792.64
50%	1,012.37	1,278.52	1,351.75
75%	1,848.86	2,128.36	2,262.46
Max	9,557.12	10,442.04	8,172.03

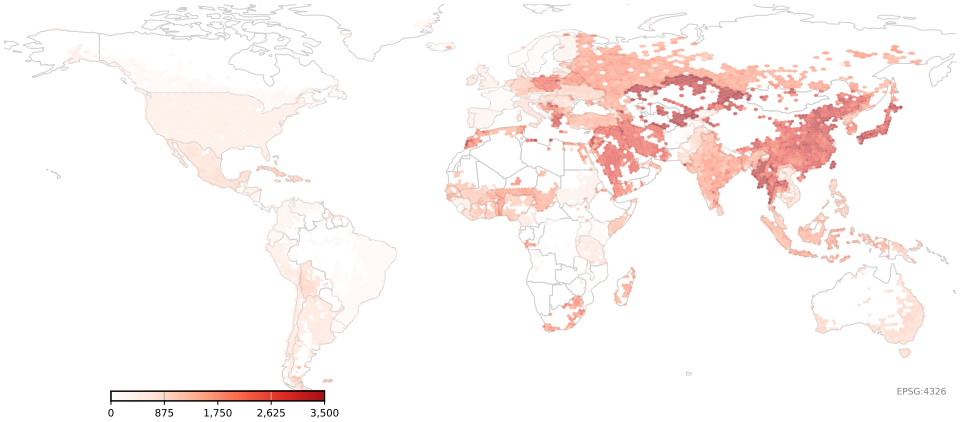
The environmental cost index in Table 5.11 shows considerable variation across cells. For GPT-4o the mean index value is 1,293 with a median of 1,012 and a maximum of 9,557. The distribution is right-skewed, indicating that a smaller number of cells contribute disproportionately high values. DeepSeek-LLM-7B and Mistral-7B show higher mean index values of 1,504 and 1,665 respectively, consistent with their lower overall token efficiency levels observed in earlier sections. Values above 3,500 are present but represent only the upper tail of the distribution.

At the country level, the results are shown in Table 5.12. Turkmenistan, Myanmar, Uzbekistan, Taiwan, and Kazakhstan show the highest mean environmental cost index values across all three models. These countries combine low token efficiency with high or moderate carbon intensity grids. Several countries in Sub-Saharan Africa, including the Democratic Republic of Congo and Ethiopia, show low environmental cost index values despite moderate token efficiency, reflecting their reliance on hydroelectric power and correspondingly low national carbon intensity values. Costa Rica, Paraguay, and Nepal show the lowest index values in the dataset, combining relatively high token efficiency with near-zero carbon intensity electricity grids.

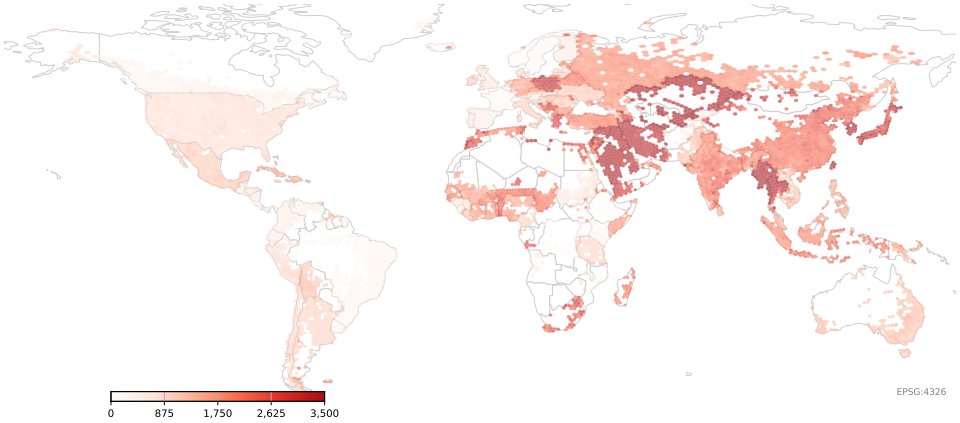
The spatial distribution of the environmental cost index is shown in Figure 5.7 for all three models. The highest values are concentrated in East Asia, Central Asia, and parts of the Middle East. The Americas and Western Europe show consistently lower index values across all three models. The spatial pattern is largely consistent across tokenizer architectures, which aligns with the broader finding that the relative ordering of regions remains stable across models.

Table 5.12: Top 10 and bottom 10 countries by mean environmental cost index. Mean token ratio represents the average across all three models. The index is computed as the product of token inefficiency and national carbon intensity of electricity (gCO2 per kWh, Our World in Data 2022). Countries with an index value of 0 are excluded from the bottom 10.

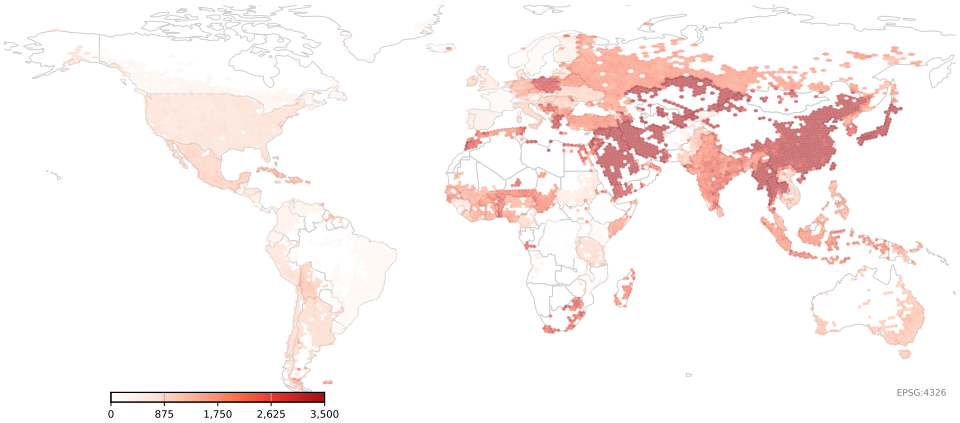
Country	**Cells**	**CO2**	**Token Ratio**	**GPT-4o**	**DeepSeek**	**Mistral**
Top 10 — highest environmental cost index						
Turkmenistan	13	1,306.14	0.274	4,171.40	5,300.26	5,715.00
Myanmar	50	563.87	0.152	4,063.99	5,306.76	4,813.07
Uzbekistan	22	1,121.35	0.273	3,655.39	4,390.60	4,605.67
Taiwan	5	639.23	0.172	3,446.33	3,438.30	4,488.44
Kazakhstan	102	831.01	0.233	3,432.72	3,725.76	3,809.97
Japan	49	519.60	0.159	3,091.80	3,175.39	3,849.31
Libya	5	830.53	0.225	2,988.94	4,239.92	4,398.92
Mongolia	21	814.19	0.275	2,832.27	3,073.94	3,261.40
China	512	586.75	0.210	2,823.42	2,365.60	3,628.86
Israel	5	595.86	0.199	2,731.46	3,144.08	3,304.78
Bottom 10 — lowest environmental cost index						
Costa Rica	7	24.12	0.458	44.49	56.06	60.26
Paraguay	14	24.71	0.416	50.02	63.52	68.42
Lesotho	1	20.83	0.352	50.44	63.70	65.71
Nepal	11	23.36	0.376	56.31	71.85	75.15
Ethiopia	63	23.55	0.350	56.82	74.37	76.65
Bhutan	3	24.14	0.363	57.47	69.36	75.78
Albania	2	24.39	0.299	69.76	87.61	91.53
DR Congo	59	27.40	0.324	72.00	93.82	96.33
Switzerland	4	37.26	0.396	80.47	103.48	106.51
Norway	46	29.64	0.247	103.50	131.41	132.47



(a) GPT-4o



(b) DeepSeek-LLM-7B



(c) Mistral-7B

Figure 5.7: Relative environmental cost index per H3 cell for the three evaluated models. Values above 3,500 are clipped.

5.6.4 Internet Penetration

Internet penetration was selected as a fourth comparison layer to examine whether the spatial distribution of token efficiency is associated with the distribution of digital infrastructure. This comparison is motivated by the argument that token efficiency may overlap with existing patterns of digital inequality, where regions with limited internet access also face structural disadvantages in how their languages are represented within large language models. Internet penetration data were obtained from the World Bank, sourced from the International Telecommunication Union, and represent the share of individuals using the internet as a percentage of the national population for the year 2022 (World Bank, 2022).

The association between token efficiency and internet penetration was examined using Spearman rank correlation. As shown in Table 5.13, all three models show a weak but statistically highly significant positive relationship ($p < 0.001$), with coefficients of 0.216 for GPT-4o, 0.178 for DeepSeek-LLM-7B, and 0.238 for Mistral-7B. This indicates that cells located in countries with higher internet penetration rates tend to show slightly higher token efficiency values. The correlation values are consistent in direction across all three models but remain weak.

Table 5.13: Spearman rank correlation of mean token ratios per H3 cell against national internet penetration rate (%, Our World in Data 2022).

Model	**Spearman r (Internet Penetration)**
GPT-4o	0.216
DeepSeek-LLM-7B	0.178
Mistral-7B	0.238

6 Discussion

6.1 Spatial Structure of Token Efficiency

The results show that token efficiency is not randomly distributed across space, but follows a clear and consistent spatial structure. This pattern is visible across all levels of analysis, ranging from individual H3 cells to macro regions and countries. Global Moran's I values indicate strong spatial autocorrelation, and local clustering analysis further confirms that high and low efficiency values are geographically concentrated rather than randomly distributed.

The coherence of these spatial patterns is of particular importance. High efficiency cells form large, continuous regions rather than appearing as isolated observations, and low efficiency cells follow a similar structure. These clusters extend across national borders and cover large geographic areas, indicating that token efficiency behaves as a spatially structured variable. Neighboring cells tend to show similar efficiency values, which suggests that local variation alone cannot account for the observed distribution.

At the same time, the spatial structure is not uniform in its form. Some regions exhibit relatively homogeneous efficiency levels, with limited variation between neighboring cells. Other regions show lower efficiency values but remain equally compact, forming large zones of consistently low performance. In contrast, regions such as Southeast Asia and South Asia display stronger internal variation, where higher and lower efficiency cells occur in close proximity. These regions are also characterized by high linguistic diversity, which may contribute to the more fragmented spatial structure observed there.

This distinction between more homogeneous and more heterogeneous regions indicates that spatial clustering does not follow a single global pattern. The transition between stable clusters and more fragmented distributions appears gradual rather than abrupt, which is consistent with the geographic distribution of language families across space. These patterns are consistently visible across models and are supported by both global and local spatial statistics.

The spatial structure of token efficiency does not exist independently of the underlying linguistic context. The distribution of languages and writing systems across the globe is itself spatially structured, and this distribution is reflected in the observed efficiency patterns. This suggests that the observed spatial patterns are closely linked to the geographic distribution of language rather than being driven by spatial proximity alone (Petrov et al., 2023; Rahman et al., 2024).

6.2 Stability Across Tokenizer Architectures

The spatial patterns identified across models show a high degree of consistency. Although GPT-4o, DeepSeek-LLM-7B, and Mistral-7B differ in their average efficiency levels, these differences do not lead to changes in the overall spatial structure. All three models produce highly similar geographic distributions of token efficiency.

This stability is visible across multiple levels of analysis. At the level of individual H3 cells, efficiency values are strongly correlated between models, indicating that cells with higher or lower efficiency in one model tend to show similar relative values in the others. At the level of macro regions and countries, the relative ordering of regions remains largely unchanged, which suggests that the global ranking of efficiency is consistent across models.

The rank correlations further support this observation. The same countries appear at the top and bottom of the efficiency rankings across all three models. While the absolute token ratios differ between models, their relative positions within the rankings remain largely unchanged.

These findings suggest that the spatial organization of token efficiency is not specific to a single tokenizer architecture. This pattern persists across independently developed systems from different centers of AI development, which indicates that it reflects something more structural than individual design choices. Within the class of subword based tokenization methods represented by the three analyzed models, differences between tokenizers are mainly associated with the magnitude of efficiency rather than its spatial distribution. Whether fundamentally different approaches such as character level or morphology aware tokenization would produce different spatial distributions remains an open question. This connects the empirical results to the broader question of global inequality in LLM performance, which is addressed in the following sections.

6.3 Token Efficiency and External Geographic Indicators

The three external data layers provide additional context for the spatial patterns identified in the preceding sections. Each layer is examined separately. No causal relationships are assumed.

The writing system comparison shows the strongest association with token efficiency. Latin Basic achieves the highest median efficiency values across all three models, while script groups such as Myanmar, Greek, and Hangul are concentrated at the lower end of the distribution. This pattern is consistent across models and spatial scales, which suggests that the relationship between script type and token efficiency is not model specific. Subword tokenization strategies rely on frequent substring reuse, a property that aligns well with alphabetic scripts but may be less suited to logographic or syllabic writing systems (Rahman et al., 2024; Sennrich et al., 2016). The consistency of this pattern across independently developed tokenizers indicates that it reflects structural differences in how writing systems interact with current segmentation strategies rather than individual training decisions.

The population comparison shows a weak negative association between population density and token efficiency. Cells with higher population values tend to show slightly lower efficiency values across all three models. While population density shows a measurable relationship with token efficiency, the results indicate that writing systems provide a more consistent explanation for the observed global patterns. This suggests that the observed spatial structure is less a function of where people live, and more a function of how language is written.

The combined analysis of writing system and population further contextualizes these findings. Script groups associated with the lowest token efficiency values are not concentrated in sparsely populated areas. CJ cells show the highest mean population sum per cell at 2,429,204, combined with a mean token ratio of 0.206 for GPT-4o. Hangul and Thai cells show similarly high mean population values of 1,604,289 and 1,722,324 respectively, combined with consistently low token efficiency across all three models. This indicates that the tokenization disadvantage associated with these writing systems is not limited to marginal or remote areas, but is concentrated in densely populated regions where large numbers of people interact with large language models. In contrast, Cyrillic cells show the lowest mean population sum at 328,004, despite covering a large number of cells across Russia and Central Asia. This suggests that the geographic extent of a script group does not determine its population exposure to tokenization inefficiency. The script groups that combine low efficiency with high population values represent the most direct overlap between linguistic disadvantage and demographic scale identified in this thesis.

The carbon intensity comparison indicates that regions with low token efficiency are in several cases also located in countries with higher carbon intensity electricity grids. The relative environmental cost index shows the highest values in East Asia and Central Asia across all three models. Countries such as Turkmenistan, Myanmar, and Kazakhstan combine low token efficiency with high national carbon intensity, while countries such as Costa Rica, Paraguay, and Nepal show low index values due to their reliance on hydroelectric power. This association is further supported by moderate negative Spearman correlations between token efficiency and national carbon intensity, ranging from -0.332 for DeepSeek-LLM-7B to -0.391 for Mistral-7B, all statistically highly significant ($p < 0.001$). These results are descriptive and the index does not represent an absolute measure of energy consumption or emissions. They suggest however that the spatial distribution of token efficiency may overlap with existing patterns of environmental cost in ways that are worth examining in future research.

The internet penetration comparison shows a weak positive association between national internet access rates and token efficiency. Cells located in countries with higher internet penetration rates tend to show slightly higher token efficiency values across all three models, with Spearman correlations ranging from 0.178 for DeepSeek-LLM-7B to 0.238 for Mistral-7B, all statistically highly significant ($p < 0.001$). The magnitude of this relationship is limited, and considerable variation exists within similar internet penetration ranges. This suggests that digital infrastructure access alone does not determine token efficiency. The association is nonetheless consistent across models and aligns with the broader argument that regions with weaker digital infrastructure may face compounded disadvantages, both in terms of internet access and in how their languages are represented within large language models (Graham et al., 2015; Peng, 2024).

6.4 Structural Sources of Spatial Inequality in Token Efficiency

The spatial patterns identified in this thesis raise the question of why certain regions perform differently than others. While the analysis does not directly measure linguistic or orthographic properties, the consistency of the observed patterns across models, spatial scales, and aggregation levels allows for an interpretation of potential structural sources underlying differences in token efficiency.

Regions associated with alphabet-based writing systems, especially Latin scripts, consistently show high token efficiency across all models. North America, large parts of Western Europe, Latin America, and Oceania form stable clusters of high efficiency,

with relatively low internal variation. These observations suggest that subword tokenization strategies align well with the structure of alphabet-based scripts, where recurring character sequences and shared morphemes can be reused across words and place names. Previous work has shown that such properties are well suited for BPE and Unigram-based tokenization approaches, which rely on frequent substring reuse to produce compact representations (Rahman et al., 2024; Sennrich et al., 2016).

In contrast, regions dominated by non-Latin writing systems consistently show lower efficiency values. These regions also form large spatial clusters of low efficiency, particularly in East Asia, Southeast Asia, and parts of Eastern Europe. The consistency of these patterns across independently developed tokenizers suggests that they are not model specific, but reflect structural differences in how these writing systems interact with tokenization processes.

DeepSeek-LLM-7B shows a visible but limited efficiency advantage in CJ script regions relative to GPT-4o and Mistral-7B. This difference is explainable with its development context and exposure to Mandarin training data. But still, CJ script regions remain among the lowest efficiency areas for DeepSeek as well, and the overall spatial hierarchy is not altered by this advantage. The persistence of low CJ efficiency across all three models suggests that the observed inefficiencies reflect the interaction between current subword segmentation strategies and character-based writing systems more broadly. Training data composition alone does not appear to be the primary driver of these patterns.

Using place names as textual input further enhances these effects. Place names often preserve historical orthography and local naming conventions that diverge from standardized language corpora, as well as compound character structures. In regions that use non-alphabetic scripts, place names may have little in common with the substrings learned during tokenizer training, resulting in consistently lower efficiency. This might explain why the spatial patterns in this thesis are so pronounced and why low efficiency appears as a coherent regional characteristic rather than as isolated local variation.

Finally, it should be noted that this analysis does not assign efficiency values to individual languages or scripts directly. Spatial units may contain multiple linguistic contexts, transliterations, or hybrid naming conventions, and no explicit language identification was performed. Orthography therefore serves as an interpretive framework rather than a directly measured variable. Nevertheless, the consistency of the observed patterns across regions and models suggests that these structural interactions are strong enough to shape global spatial patterns of token efficiency.

6.5 Implications for Representation, Cost, and Access

The results show that token efficiency is uneven, spatially structured, and consistent across models and spatial scales. These patterns are associated with how tokenization relates to representation, computational cost, and model evaluation.

A consistent pattern across all three models is the similarity in spatial distribution of efficiency values. Regions with higher efficiency in one model also show higher efficiency in the others. The same holds for regions with lower efficiency. The models differ in overall efficiency levels, but the spatial hierarchy remains stable. This suggests that within subword-based tokenization, model design is associated with differences in magnitude, not with differences in spatial distribution.

The observed patterns are spatially autocorrelated and form large, continuous regions of similar efficiency values. These clusters remain visible when aggregating results from H3 cells to larger spatial units such as regions or countries. This indicates that the distribution of token efficiency is not dependent on the level of spatial aggregation.

The similarity across independently developed tokenizers suggests that token efficiency is linked to structural properties of the input text. Writing systems are unevenly distributed across space, and efficiency values follow these distributions. Regions dominated by non-Latin scripts show lower efficiency and form spatial clusters of similar values. Regions using Latin-based scripts show higher efficiency and more compact token representations.

These findings suggest that tokenization shapes how text is represented before model inference. Token efficiency determines how many tokens are required to encode the same content, which is directly associated with the number of computational steps during processing. Since the spatial patterns are similar across models, these differences are not reduced by changes in model architecture.

This can be interpreted as an indication that part of the variation in computational cost is introduced at the preprocessing stage. Languages and regions associated with lower efficiency require more tokens for equivalent input, which is associated with higher resource use during inference. Given that commercial systems often operate on a per-token basis, this may also be associated with differences in monetary cost.

From a representation perspective, the results suggest that languages are not encoded with the same level of compactness across space. Regions with lower efficiency are represented through longer token sequences, which may affect how information is processed within the model. This can be interpreted as a difference in how linguistic structures are translated into computational units.

In the context of model evaluation, these findings suggest that global average metrics may mask spatial variation in token efficiency. Models that show similar overall performance may still differ in how efficiency is distributed across regions. This indicates that spatial patterns should be considered when comparing models in multilingual settings.

Putting these findings together suggests a relationship between token efficiency and the broader questions raised in the theoretical framework. Tokenization is not a neutral preprocessing step, but a mechanism that shapes how language is represented within large language models. The observed patterns suggest that tokenization may reflect existing global linguistic structures, contributing to spatially uneven representation within large language models. Recognizing this shifts the focus from treating token efficiency as a purely technical metric to understanding it as a structural property of language models that shapes the economic and *representational geography*.

6.6 Limitations

This thesis offers a comprehensive analysis of the spatial structure of token efficiency, but several limitations define the interpretation of the results.

The analysis relies on place names obtained from volunteered geographic information. This approach enables global coverage and a consistent spatial framework, but also means that the input text reflects naming conventions rather than natural language use. Place names are often shorter and highly structured, shaped by historical, administrative, and cultural factors. This means that the measured token efficiency captures how tokenizers process geographic naming systems rather than complex and longer natural language.

Token efficiency is not linked to language labels or a quality score. The analysis takes place at the spatial unit level and does not assign efficiency values to individual languages. Each cell may contain place names from multiple linguistic contexts, and there is no attempt to identify the dominant language within a cell. This limits the ability to draw direct conclusions about the performance of tokenizers for specific languages.

The three analyzed models all rely on subword segmentation as their core mechanism. The stability observed across models therefore applies to this class of tokenization methods rather than to tokenization in general. Fundamentally different approaches such as character level or morphology aware tokenization were not examined, and the findings cannot be extended to those methods without further analysis.

The relative environmental cost index introduced in section 5.6.3 combines two independently measured variables into a single comparable value. This index is exploratory and

does not represent an absolute measure of energy consumption or emissions. The choice to multiply token inefficiency by national carbon intensity is a simplification that does not account for variation in data center locations, energy sources at the infrastructure level, or actual inference costs. Results from this index should be interpreted with caution.

The three external data layers, internet penetration, carbon intensity, and dominant writing system, are not all represented at the same spatial resolution. While internet penetration and carbon intensity are assigned at the national level and matched to H3 cells based on country boundaries, the dominant writing system is derived directly from the place names within each cell. Population data is aggregated at the H3 level from a global raster dataset. This means that some variables capture local variation, while others do not. As a result, within-country differences are only partially represented, and cells in large or diverse countries may not fully reflect local conditions. Internet penetration and carbon intensity data are both from 2022 and represent a static snapshot that may not reflect current conditions in all regions

The spatial resolution of the analysis was limited by computational feasibility, as the H3 resolution was set at level 3. This enabled a global analysis with reasonable runtime, but a higher resolution could have captured finer grained variation. Resolution 3 was already requiring dozens of hours per region, making a higher resolution unfeasible within the scope of this thesis.

Finally, the study period is static. Data were collected at one point in time, creating a snapshot of the digital linguistic landscape of mid 2025. Changes over time, such as tokenizer updates or shifts in OSM coverage, are outside the scope of this thesis. This would allow for a closer link between spatial patterns and actual language use.

7 Conclusion

This thesis examined token efficiency as a geospatial phenomenon with global implications. Starting from the observation that tokenization has been studied primarily as a linguistic or technical process, this work introduced a spatial perspective by mapping token efficiency across the world using a global H3 grid, place name based textual input from OpenStreetMap, and three tokenizer architectures from different centers of AI development.

The first research question asked how much token efficiency varies across different regions worldwide. The results show that the variation is substantial and geographically structured. Token efficiency does not appear randomly distributed across space. Instead, it forms large, coherent zones of consistently high or low performance that persist across spatial scales and aggregation levels. High efficiency regions are concentrated in areas dominated by Latin script languages, particularly English and Spanish speaking parts of the world, while low efficiency regions are found across East Asia, Southeast Asia, the Middle East, and parts of Eastern Europe. The strength of this spatial structure was confirmed statistically through global Moran's I values above 0.92 for all three models, and local clustering analysis revealed that these patterns emerge through large continuous zones rather than isolated local effects.

The second research question asked how different tokenizers perform globally and whether their spatial patterns differ. The results show that GPT-4o, DeepSeek-LLM-7B, and Mistral-7B reproduce nearly identical spatial hierarchies. Cross model Pearson correlations exceed 0.96 at the cell level, and Spearman rank correlations at the country level reach up to 0.988. The same countries appear at the top and bottom of the efficiency rankings across all three models. The results show that global differences in token efficiency are not primarily determined by tokenizer design, but by structural properties of writing systems and their spatial distribution. Contrary to initial expectations, differences between tokenizer architectures do not substantially alter the global distribution of efficiency, but instead preserve a shared spatial structure. It should be noted that all three models rely on subword segmentation as their core mechanism. The consistency observed here therefore applies to this class of tokenization methods rather than to tokenization in general.

The third research question asked what the observed spatial patterns suggest about digital inequality and the environmental implications of tokenization. The four external data layers provide additional context here. The writing system comparison shows the strongest association with token efficiency, with Latin Basic consistently achieving the highest efficiency values and script groups such as Myanmar, Greek, and Hangul concentrated at the lower end across all models. The population comparison shows a weak negative association, suggesting that population distribution alone does not account for the observed patterns. The internet penetration comparison shows a weak positive association, with cells in countries with higher internet access rates tending toward slightly higher token efficiency values, a pattern that is consistent across all three models. The carbon intensity comparison indicates that regions with low token efficiency are in several cases also located in countries with higher carbon intensity electricity grids, with moderate negative Spearman correlations between token efficiency and national carbon intensity across all three models. These results are descriptive and do not establish causal relationships, but they suggest that the spatial distribution of token efficiency overlaps with existing patterns of digital and environmental inequality in ways that are worth examining further. A full answer to RQ3 would require more direct measures of digital infrastructure and energy consumption, and this remains an area for future research.

What the results demonstrate clearly is that inefficiency is embedded at the tokenization stage, before any model inference begins, meaning it is difficult to correct through improvements to downstream model behavior alone. Users in persistently low efficiency regions face higher computational and economic costs for equivalent tasks, and this disparity is consistent across all three analyzed architectures. Tokenization, in this sense, is not a neutral step. It distributes the cost of language processing unevenly across geographic space, and that distribution follows patterns that are worth examining alongside broader questions of who benefits from and who is burdened by current AI systems.

This thesis introduces a methodology that is reproducible and transferable. The combination of a global discrete grid, open geographic data and multiple tokenizer comparisons provides a framework that can be extended to higher spatial resolutions, additional models, or alternative text inputs. A natural next step would be to replace place names with standardized natural language samples in locally dominant languages, which would bring the analysis closer to actual language use and allow for a more direct connection between spatial patterns and specific linguistic structures. Linking token efficiency directly to sustainability metrics such as energy consumption per token, or to socioeconomic indicators such as internet access or income, would also allow the interpretive arguments made here to be tested more rigorously.

The contribution of this thesis is not to solve the problem of unequal tokenization, but

to make it visible as a geographic phenomenon. By showing *where* inefficiency is located, how stable it is, and that it cannot be attributed to any single model, this work establishes token efficiency as an additional spatial baseline for evaluating fairness in large language models. The geography of tokenization is not a side effect of how these systems are built. It is part of their structure.

8 Outlook

This thesis grew out of a genuine interest in combining two fields that do not often speak to each other directly. Studying how artificial intelligence and geography intersect was one of the most engaging parts of my Master's, and this project felt like a natural expression of that interest. Geography has always been concerned with how phenomena are distributed across space and why those distributions are not random. Applying that logic to something as technical and *invisible* as tokenization felt like an unusual but worthwhile direction. Mapping token efficiency onto the surface of the Earth turned out to be more revealing than I initially expected, and working at the boundary between computational linguistics and spatial data science was something I found genuinely motivating throughout the entire process. The finding that surprised me most was the consistency across tokenizer architectures. Going into this project I was convinced that each model would behave differently in its own linguistic domain. It seemed logical that DeepSeek-LLM-7B, developed with a strong orientation toward Mandarin and trained on large amounts of non Latin script data, would handle those languages as efficiently as GPT-4o handles English. Seeing that this is simply not the case was unexpected. No matter which model was analyzed, the spatial hierarchy of efficiency remained almost identical. The same regions performed well, the same regions performed poorly, and the differences between models amounted to shifts in level rather than changes in structure. That result forced me to reconsider where the source of these inefficiencies actually lies. It is not primarily a question of which company built the model or which language they prioritized. It runs deeper than that, into the segmentation strategies themselves and how they interact with different writing systems. I had not anticipated that conclusion going in, and it is the part of this thesis I find most worth thinking about.

What this thesis reinforced for me on a broader level is that AI does not exist outside of the world it was built in. It reflects the same imbalances that already exist: in data, in infrastructure, in which languages and places are treated as standard and which are treated as exceptions. Working through this analysis made visible just how disadvantaged large parts of the world are within these systems, in ways that are quiet and technical enough to go largely unnoticed in everyday use. Someone using a large language model in rural Myanmar or rural Georgia is paying more, computationally and often financially,

for the same task as someone using it in the United States, not because of any deliberate decision, but because of how the system was built from the ground up. That is not a small thing. There are still many structural problems that need to be addressed before these technologies distribute their benefits more evenly, and this thesis is a small contribution toward making those problems visible and measurable.

If I could extend this work further, I would go deeper into the relationship between individual language structure and tokenization outcomes. Understanding not just where inefficiency is located spatially, but why specific languages are tokenized so poorly and what the internal linguistic mechanisms behind that are, feels like the most natural continuation of this work. The spatial patterns identified in this thesis raise a set of questions that a more linguistically focused analysis could begin to answer. That connection between place, language and computational representation is something I would have liked to explore further within this thesis, and it remains the direction I find most worth pursuing.

Code Availability

The code used in this thesis is available at: https://github.com/flocoloco4/Master-Thesis-Venier

The repository contains all scripts required to reproduce the analysis and generate the results presented in this thesis.

Bibliography

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A., & Tsvetkov, Y. (2023). Do all languages cost the same? tokenization in the era of commercial language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 9904–9923. https://doi.org/10.18653/v1/2023.emnlp-main.614

Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Schulze Buschhoff, J., Jain, C., Weber, A. A., Jurkschat, L., Abdelwahab, H., John, C., Ortiz Suarez, P., Ostendorff, M., Weinbach, S., Sifa, R., … Flores-Herr, N. (2024). Tokenizer choice for LLM training: Negligible or crucial? *Findings of the Association for Computational Linguistics: NAACL 2024*, 3907–3924. https://doi.org/10.18653/v1/2024.findings-naacl.247

Ballatore, A., Graham, M., & Sen, S. (2017). Digital hegemonies: The localness of search engine results. *Annals of the American Association of Geographers*, *107*(5), 1194–1215. https://doi.org/10.1080/24694452.2017.1308240

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Berrill, T., Mah, J., Moran, A., et al. (2022). H3 hexagonal spatial indexing for environmental and spatial modelling.

Birch, C. P. D., Oom, S. P., & Beecham, J. A. (2007). Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. *Ecological Modelling*, *206*, 347–359.

Brodsky, I. (2018). H3: Uber's hexagonal hierarchical spatial index [Accessed 2025]. https://www.uber.com/blog/h3/

DeepSeekAI. (2024). Deepseek coder: Open multilingual large language model for code and natural language. *arXiv preprint arXiv:2401.14196*. https://arxiv.org/abs/2401.14196

Ember & International Energy Agency. (2022). Carbon Intensity of Electricity Generation [Accessed 2025].

European Commission, Joint Research Centre. (2023). GHS Population Grid, Epoch 2025, Global, Resolution 1km, R2023A [Accessed 2025].

Goldman, O., Caciularu, A., Eyal, M., Cao, K., Szpektor, I., & Tsarfaty, R. (2024). Unpacking tokenization: Evaluating text compression and its correlation with model performance. *Findings of the Association for Computational Linguistics: ACL 2024*, 2274–2286. https://doi.org/10.18653/v1/2024.findings-acl.134

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221.

Goodchild, M. F., Kimerling, A. J., & Sahr, K. (2020). Discrete global grid systems for earth system science. *International Journal of Digital Earth*, *13*(6), 732–753. https://doi.org/10.1080/17538947.2019.1701558

Goodchild, M. F., & Longley, P. A. (2021). Geocomputation and GIScience [Available at: https://discovery.ucl.ac.uk/id/eprint/10053533/]. In D. Richardson et al. (Eds.), *International encyclopedia of geography*. John Wiley & Sons.

Graham, M., De Sabbata, S., & Zook, M. A. (2015). Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo: Geography and Environment*, *2*(1), 88–105. https://doi.org/10.1002/geo2.8

Janowicz, K. (2023). Philosophical foundations of GeoAI: Exploring sustainability, diversity, and bias in GeoAI and spatial data science. https://arxiv.org/abs/2304.06508

Janowicz, K., Mai, G., Huang, W., Zhu, R., Lao, N., & Cai, L. (2025). Geofm: How will geo-foundation models reshape spatial data science and geoai? *International Journal of Geographical Information Science*, *39*(9), 1849–1865. https://doi.org/10.1080/13658816.2025.2543038

Jiang, A., Lambert, M., Le Scao, T., et al. (2024). Mistral models: Efficient open foundation models. *arXiv preprint arXiv:2401.04088*. https://arxiv.org/abs/2401.04088

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the nlp world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Li, P., Yang, J., Islam, M. S., & Ren, S. (2023). Making AI less thirsty: Uncovering and addressing the secret water footprint of AI models. https://doi.org/10.48550/arXiv.2304.03271

Liu, Z., Janowicz, K., Majic, I., Shi, M., Fortacz, A., Karimi, M., Mai, G., & Currier, K. (2025). Operationalizing geographic diversity for the evaluation of ai-generated content. *Transactions in GIS*, *29*(3), e70057.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., & Tan, S. (2021). Between words and characters: A brief

history of open-vocabulary modeling and tokenization in NLP. https://arxiv.org/abs/2112.10508

Mor, N. (2025). It's a global village (if you speak the right language): On language models, digital sidelining, and participation. *Wisconsin International Law Journal*, *42*, 329.

Mosa, H., Saleh, A., & Al-Badarneh, A. (2025). Performance comparison of spatial data indexing using distributed systems, 327–333. https://doi.org/10.1109/ICTCS65341.2025.10989398

OpenAI. (2023). *Gpt-4 technical report* (tech. rep.) (Describes model architecture, tokenizer, and training scope of GPT-4 family). OpenAI. https://cdn.openai.com/papers/gpt-4.pdf

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*. https://doi.org/10.48550/arXiv.2104.10350

Peng, L. (2024). *Artificial divides: Global AI access disparities and constructions of new digital realities* [Master's thesis]. University of Washington [Department of Geography. Committee: Bo Zhao, Mia Bennett].

Petrov, A., La Malfa, E., Torr, P. H. S., & Bibi, A. (2023). Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, *36*, 36963–36990. https://arxiv.org/abs/2305.15425

Rahman, A., Bowlin, G., Mohanty, B., & McGunigal, S. (2024). Towards linguistically-aware and language-independent tokenization for large language models (llms). https://arxiv.org/abs/2410.03568

Santorelli, M., Catullo, D., & Palladino, M. (2024). Ai and the reduction of social inequalities in a linguistic perspective. *Proceedings of the Conference on Quality Evaluation in Governance (CECG)*, 144–154. https://doi.org/10.36004/nier.cecg.II.2024.18.14

Schmidt, F., Kauschke, C., & Schulte im Walde, S. (2024). Beyond compression: Evaluating tokenizers for language models. *Transactions of the Association for Computational Linguistics*, *12*, 1–18. https://doi.org/10.1162/tacl_a_00633

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Shahid, F., Elswah, M., & Vashistha, A. (2025). Think outside the data: Colonial biases and systemic issues in automated moderation pipelines for low-resource languages.

Shi, M., Janowicz, K., Verstegen, J., Currier, K., Wiedemann, N., Mai, G., Majic, I., Liu, Z., & Zhu, R. (2025). Geography for ai sustainability and sustainability for geoai. *Cartography and Geographic Information Science*, *52*(4), 331–349. https://doi.org/10.1080/15230406.2025.2479796

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. https://arxiv.org/abs/2302.13971

Velayuthan, M., & Sarveswaran, K. (2025). Egalitarian language representation in language models: It all begins with tokenizers. *Proceedings of the 31st International Conference on Computational Linguistics*, 5987–5996. https://aclanthology.org/2025.coling-main.400/

Wang, Z., Wu, N., Cao, Q., Xia, J., Liu, Z., Xie, Y., Nambi, A., Ganu, T., Lao, N., Liu, N., & Mai, G. (2025). Geobs: Information-theoretic quantification of geographic bias in ai models. *arXiv preprint arXiv:2509.23482*. https://doi.org/10.48550/arXiv.2509.23482

Wilhelm, P., Wittkopp, T., & Kao, O. (2025). Beyond test-time compute strategies: Advocating energy-per-token in LLM inference. *Proceedings of the 5th Workshop on Machine Learning and Systems (EuroMLSys 2025)*, 1–8. https://doi.org/10.1145/3721146.3721953

World Bank. (2022). Individuals Using the Internet (% of Population), Indicator IT.NET.USER.ZS [Data sourced from the International Telecommunication Union. Accessed 2025].