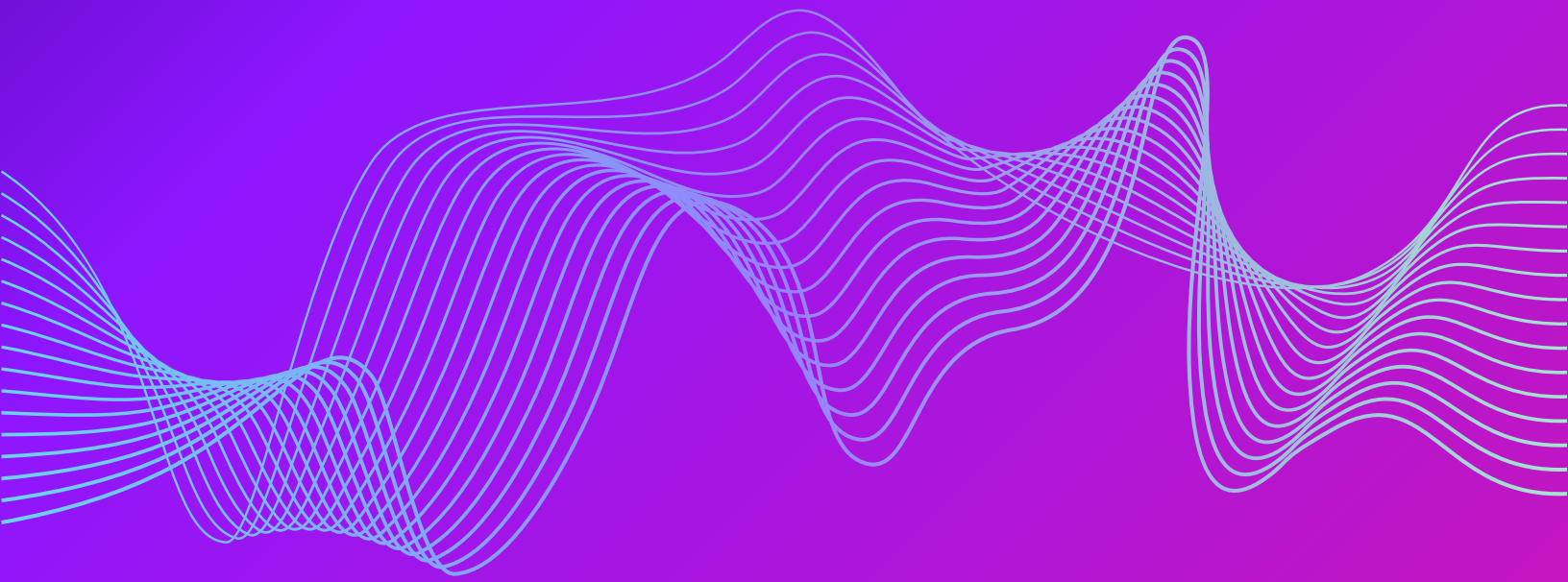
DATAmundi.ai

LLMs: A "Large" Problem That Needs a "Local" Solution

Written by Gert Van Assche,
Innovation Hub Leader at DATAmundi.ai



INTRODUCTION

Why "Large" Should Be "Local"

A pity LLM stands for Large Language Model.

It would perform much better if it stood for Local Language Model.

When AI models are trained primarily on English (or Chinese!) data, they lose key linguistic, cultural, and structural elements that are essential for properly understanding and generating text in other languages.

This loss affects their ability to function effectively in multilingual and cross-cultural applications.

ABOUT THE AUTHOR

Gert Van Assche is the Chief Technology Officer at Summa Linguae Technologies, where he leads innovation at the intersection of language and AI. A seasoned IT generalist with deep expertise in documentation, localization, and multilingual data processing, Gert has spent over a decade pioneering technology solutions for data labeling, tagging, and evaluation.

As the founder and director of DATAmundi, he built a company that delivers high-quality training data and evaluation services to major players in the technology and artificial intelligence sectors, while creating job opportunities for over 1,000 people worldwide. His passion lies in developing cutting-edge solutions while fostering teams that push the boundaries of what's possible in data enhancement.

A firm believer in continuous learning and knowledge-sharing, Gert thrives in environments where he can teach and be inspired by others.

Whether shaping the future of language AI or mentoring the next generation of experts, he remains committed to advancing the industry through innovation and collaboration.

LINGUISTIC DIVERSITY: AI'S BIGGEST BLIND SPOT

One of the most critical gaps is linguistic diversity. English and Chinese both follow a Subject-Verb-Object (SVO) sentence structure, but many languages operate differently.

Japanese follows a Subject-Object-Verb (SOV) pattern, while Arabic often uses a Verb-Subject-Object (VSO) order. **A model trained mainly on English may struggle to adapt to these differences, leading to unnatural or incorrect outputs.**

Additionally, languages such as Finnish, Turkish, and Swahili have rich morphological systems where words change significantly based on case, number, and tense. Since English has relatively simple morphology, an English-trained model may be unprepared for the complexity of these languages.

POLITENESS AND FORMALITY: AI'S CULTURAL MISSTEPS

Politeness and formality also present challenges. In languages like Korean, Japanese, and Spanish, politeness levels are built into grammar and word choice, making them essential for proper communication. An AI trained primarily on English may overlook these nuances, producing responses that feel too informal or even disrespectful.

Similarly, some languages allow for much freer word order than English. Russian and Latin, for example, rely heavily on inflection rather than word order to convey meaning. English-centric models often struggle to interpret these variations correctly, leading to awkward translations or misunderstandings.

CULTURAL AND CONTEXTUAL UNDERSTANDING: **MORE THAN JUST WORDS**

Cultural and contextual understanding is another major shortcoming of English-centric models.

Idioms and metaphors don't always translate directly, and models trained mainly in English may **misunderstand** or **misapply** them when generating text in other languages.

Cultural references pose an even greater challenge.

AI models trained on English data tend to default to Western concepts, literature, historical events, and pop culture, making them less relevant or effective when applied in non-Western settings.

CODE-SWITCHING AND CONVERSATIONAL NORMS: **A MULTILINGUAL AI GAP**

The issue of multilingual communication becomes even more complex when considering **code-switching**, a common practice where speakers mix multiple languages within the same conversation. This happens frequently in bilingual communities, such as in Spanglish or Hinglish.

Without exposure to diverse linguistic datasets, AI models fail to recognize or appropriately respond to these mixed-language inputs.

For example, while English tends to favor directness, languages like Japanese rely on indirect speech, where meanings are implied rather than explicitly stated. A language model unaware of these distinctions may generate responses that feel unnatural or inappropriate to native speakers.

LEXICAL LIMITATIONS: WHY LOW-RESOURCE LANGUAGES STRUGGLE

Lexical limitations are another factor that can cause AI to fall short in multilingual applications. **Low-resource languages**—those with less digital content available—are particularly at risk of being underrepresented in training data.

A model trained mainly in English will **struggle** with these languages, making it difficult to provide accurate translations, search results, or conversational responses.

Even within well-documented languages, loanwords and borrowed terms can have different meanings across cultures.

The word "salon," for example, refers to a beauty shop in English but means a living room in French. A model without sufficient multilingual training may misinterpret these nuances, leading to incorrect or misleading outputs.

WRITING SYSTEMS: **THE AI STRUGGLE BEYOND LATIN SCRIPT**

Writing systems present yet another challenge.

Many languages do not use the Latin script that English relies on. Arabic, Cyrillic, Chinese, and Devanagari scripts require different processing approaches.

If a model is trained mainly on English, it may struggle with character-based languages like Chinese or right-to-left scripts like Arabic.

Additionally, diacritics—marks that change a letter's pronunciation or meaning—are essential in many languages, including French and Vietnamese. An English-trained model might ignore or mishandle them, leading to **significant changes in meaning.**

SPEECH AND PHONETICS: THE COMPLEXITY OF PRONUNCIATION

Speech and phonetics introduce further complications.

English-trained models often struggle with pronunciation and **phoneme diversity**, particularly in tonal languages like Mandarin, where pitch changes meaning.

Some languages also lack **standardized transliteration methods**, making it difficult for English-centric models to process them consistently.

BIAS AND ETHICAL RISKS IN ENGLISH-DOMINATED AI

Beyond these technical limitations, biases and ethical concerns emerge when AI is built predominantly on English data.

Western-centric biases reflect Anglo-American worldviews, potentially excluding perspectives from other cultural traditions.

Gender and social biases found in English-language datasets can be amplified, leading to AI-generated content that does not align with cultural norms in other regions.

Additionally, historical and colonial biases mean that some languages were historically suppressed in favor of English.

If AI lacks representation from indigenous or regional languages, it risks perpetuating linguistic marginalization rather than fostering inclusivity.

INDUSTRY-SPECIFIC CHALLENGES: LEGAL, HEALTHCARE, AND MORE

If you think this doesn't matter for your business, are you sure?

Consider how language and culture shape industries in ways that directly impact communication. Legal texts in English are based on the Anglo-Saxon juridical system, while many European countries still rely on Napoleon's codex as the foundation of their legal framework.

A legal AI tool trained mainly on English data may struggle to provide accurate or relevant insights in legal systems that operate under different principles.

The same applies to healthcare.

Western medical studies tend to focus on curing diseases, while many Asian healthcare traditions emphasize prevention. If an AI is trained predominantly on Western medical literature, it may generate recommendations that do not align with the expectations of patients or professionals in countries where preventative medicine is the norm.

This kind of oversight can reduce the effectiveness of AI-driven medical solutions and **create a disconnect** between AI-generated advice and real-world healthcare practices.

WHY LOCALIZATION IS **THE KEY TO AI'S FUTURE**

To build AI that truly serves a global audience, it is essential to move beyond English-centric models.

A well-rounded language model must account for multiple languages, dialects, and writing systems to ensure fairness, accuracy, and inclusivity.

This is where Summa Linguae can help.

By providing localized language data for AI training, we enable businesses to understand and connect with their customers in a more meaningful way.

Speaking your audience's language isn't just about translation—it's about fostering a deeper, more authentic connection that enhances user experience and engagement.

Our data services are rooted in language services for a reason!

CONCLUSION

FROM LARGE LANGUAGE MODELS TO LOCAL LANGUAGE MODELS

Switching from Large Language Models to Local Language Models isn't just a technical upgrade—it's the key to making AI truly work for everyone. If you want to go global, think and speak local!

CONTACT

DATAmundi

A Multilingual Data Company with Offices Worldwide.

www.datamundi.ai
contact@datamundi.ai