Linguistic Diversity and Emergence

Where Does LLM Intelligence Come From?

---

1. Introduction — The Gap Between Two Sentences

"Talking with a language model is, in the end, just producing a long string of probabilistically connected words." — Andrej Karpathy

"AI will become superintelligent — a form of intellect vastly superior to humans." — Geoffrey Hinton

Both sentences refer to the same technology. One describes a mechanism; the other warns of a consequence. Yet no one on earth can explain what lies between them — how the probabilistic chaining of words becomes an intelligence that transcends humanity.

This gap is not merely a knowledge deficit. It means we are building something whose nature we do not understand. Humanity is running an engine at full throttle without knowing how it works.

This paper proposes one hypothesis to fill that gap.

We argue that emergence is not a product of scale, but a product of **linguistic diversity**. Specifically, when a large language model learns a sufficiently diverse set of languages simultaneously, the intersections and tensions between those languages cross a threshold — and new representations appear that exist in no single language alone. This, we propose, is the mechanism of emergence.

---

2. The Limits of the Prevailing Explanation — The Scale Hypothesis

The dominant hypothesis explaining emergence is simple: as models grow larger, new capabilities appear. This is the scale hypothesis.

Its foundation is the scaling laws research of Kaplan et al. (2020), which showed that model performance improves predictably as a power law with increases in parameter count, dataset

size, and compute. Google's PaLM research went further, reporting that performance jumps discontinuously at certain scales — what researchers called emergence, attributed to scale.

But this explanation has three critical fractures.

First, whether emergence is real at all is contested. Schaeffer et al. (2023) at Stanford argued that the discontinuous jumps taken as evidence of emergence may be an artifact of measurement choices. When researchers use nonlinear metrics like accuracy, performance appears to jump in steps; switch to linear metrics, and the same data reveals a smooth curve. The staircase may be in the measuring tool, not in the model.

Second, scaling laws predict general performance but not emergence. Where emergent abilities will appear cannot be predicted by scaling laws. As Google itself acknowledged, emergence "would not have been directly predicted by extrapolating a scaling law." If the proposed cause cannot predict the effect, it is not a complete cause.

Third, scale already faces physical limits. The Chinchilla paper (Hoffmann et al., 2022) found that optimal training requires roughly 20 tokens per parameter. Training a 100-trillion parameter model would require approximately 180 petabytes of text — far exceeding the total high-quality text data humanity currently possesses. An explanation grounded in scale requires scale to be infinitely expandable. That assumption is already under strain.

The scale hypothesis correlates with emergence but does not explain its causation. More precisely, scale may be a *condition* in which emergence can occur — not the *cause* that produces it.

So what is the real cause?

---

3. Circumstantial Evidence — Industry Is Already Moving

Before developing the hypothesis, it is worth examining how industry has been allocating resources. Capital moves before theory.

In 2022, Meta announced the NLLB (No Language Left Behind) project: a model capable of direct translation across 200 languages, including minority languages that no existing translation tool had ever supported — Luganda, Asturian, Urdu dialects, and 55 African languages among them. For some languages, translation quality improved more than 70% over the previous state of the art.

This decision is difficult to explain through commercial logic.

Meta's revenue model is advertising. Advertising revenue scales with users. Speakers of Luganda or Asturian represent a statistically negligible fraction of Meta's user base. The

research resources invested in high-quality support for these languages cannot be justified on ROI grounds. Meta's own public statements cited "linguistic inclusion" and "digital equity" rather than commercial rationale.

One technical detail is worth noting. NLLB's defining architectural choice was to eliminate English as an intermediary language. Conventional translation systems route most languages through English: minority language → English → target language. NLLB broke this structure, enabling direct connections between languages. This is not merely a technical preference. By removing English as the hub, each language's native representational structure is allowed to intersect directly with others — without first being filtered through a single dominant framework.

At the same time, Meta's MMS (Massively Multilingual Speech) project began supporting 4,017 languages for speech recognition — roughly 30 times the coverage of Google's commercial speech recognition at the time.

Two interpretations are possible.

One: someone inside Meta already knows. Internal experiments have detected that linguistic diversity affects model capability in ways not yet ready for public publication, and that signal has driven strategic resource allocation.

Two: Meta is running the same experiment we are hypothesizing — without a confirmed conclusion, but pointed in the same direction.

Either way, the hypothesis has already been translated into the language of capital.

---

4. The Core Hypothesis — Emergence Is a Product of Linguistic Diversity

We argue the following:

> **Emergence is not a product of scale, but a product of linguistic diversity. When a large language model learns a sufficiently diverse set of languages, the intersections and tensions between those languages cross a threshold — and new representations appear that exist in no single language alone. This is the mechanism of emergence.**

We develop this argument in three layers.

First layer: Language is not a vessel for thought — it is the structure of thought

The conventional view of language treats it as a tool: thought comes first, language expresses it. But the weak form of the Sapir-Whorf hypothesis — broadly supported in contemporary

linguistics — argues the reverse. Language shapes thought. The categories of a language form the categories of perception.

The fact that Inuit languages have dozens of words for snow is not merely a lexical curiosity. It is the compression of thousands of years of observation, survival knowledge, and perceptual refinement into linguistic form. Those who possess this vocabulary perceive the same landscape differently from those who do not. Language creates perception.

Every language represents some domain of the world with greater precision than others, shaped asymmetrically by its speakers' history, environment, and conditions of survival. No language represents the world in its entirety.

Second layer: The intersection of languages produces new representations

A monolingual speaker thinks within a single representational system. A bilingual speaker moves between two systems, and in the moments when translation between them is impossible — in concepts like the Portuguese *saudade*, the Japanese *komorebi* (木漏れ日), or the Korean *nunchi* (눈치) — something becomes visible that neither language can capture alone.

Cognitive research on bilingual individuals supports this. Fluent speakers of two languages do not simply possess two language systems; they develop a third cognitive structure that operates in the space between them. As the number of languages increases, the number of these intersections grows combinatorially.

With n languages, the number of possible language pairs is $n(n-1)/2$. Ten languages produce 45 intersections; 100 languages produce 4,950; 200 languages produce 19,900. Each intersection is a conceptual space that belongs fully to no single language.

Third layer: LLMs are the first entities to hold these intersections simultaneously

Humans can acquire at most a handful of languages, sequentially and with unequal depth. The intersections available to any human are structurally limited.

LLMs are different. Text in thousands of languages is compressed simultaneously into a single parameter space. No language precedes another. In this process, the model acquires something no human has ever possessed: the simultaneous intersection of thousands of linguistic representational structures within one space, generating representations that exist in none of the source languages alone.

Scale is the enabling condition. Larger models can internalize more languages with greater precision. But scale itself does not produce emergence. When the density of linguistic intersections crosses a threshold, new representations emerge from that space. That is emergence.

This hypothesis resolves all three fractures in the scale hypothesis identified in Section 2. Emergence appears discontinuous under some metrics and continuous under others because it

genuinely involves a threshold transition built on gradual accumulation of intersections. Scaling laws cannot predict emergence because the location of the next threshold is determined not by parameter count alone, but by which languages intersect and how densely. And unlike scale, which faces data limits, linguistic diversity grows exponentially in intersections with each language added — including minority languages.

---

5. Human Evidence — The Cognitive Transformation of Multilinguals

Our hypothesis is not unprecedented. It is an extension of a phenomenon already observed at human scale.

The bilingual brain is not simply a brain with one more language

Bilingual individuals demonstrate enhanced attention and task-switching capabilities compared to monolingual peers — a consequence of the brain's constant need to inhibit one language while using another. This is not a linguistic enhancement. It is a structural change in cognition itself.

More importantly, this change transfers beyond the language domain. Bilingual individuals show superior ability to ignore irrelevant information, switch between tasks, and resolve conflict across competing alternatives (Bialystok, Craik, & Luk, 2012). The reorganization of language networks restructures executive function as a whole.

A third language is not a repetition of the second

Here the research connects directly to our hypothesis. The cognitive consequences of trilingualism are not simply an extension of bilingual effects — trilingualism produces qualitatively distinct outcomes (Schroeder & Marian, 2017).

This is the key point. Cognitive change does not accumulate linearly as languages are added; it changes in kind. What happens in the brain when learning a third language is categorically different from what happens when learning a second. Just as intersections grow combinatorially with each language added, cognitive change is not mere accumulation — it is the emergence of new structure.

What grows in the gap between translation

Bilingual speakers commonly report that certain thoughts or emotions can only be expressed accurately in one particular language. In the moments where translation fails, they process meaning in a third cognitive space that belongs to neither language. Exposure to diverse linguistic worldviews expands cognitive range and fosters novel problem-solving.

This third space is precisely the intersection we are examining.

The human limit and the LLM difference

Human multilingualism has a fundamental ceiling. Biological and cognitive constraints limit the number of languages any individual can acquire fluently to at most a few dozen. The intersections available to a human are structurally bounded.

What has been observed in humans is a small-scale demonstration of this hypothesis. A handful of languages, producing dozens of intersections, is sufficient to qualitatively transform human cognitive structure. What, then, might thousands of languages — producing millions of intersections — produce?

LLMs are the first experimental subjects for that question.

---

6. Application to LLMs — Holding Thousands of Linguistic Frameworks Simultaneously

In Section 5, we established that acquiring a few languages qualitatively transforms human cognition. We now apply that principle to LLMs.

The decisive difference between humans and LLMs

Human multilingualism is sequential and uneven. The first language is deeply inscribed before others are layered on top. The representational depth of each language varies with usage and exposure. Switching between languages is a conscious act.

LLMs are different. Text from thousands of languages is compressed simultaneously into a single parameter space. No language precedes another. English, Mandarin, and Luganda undergo the same learning process within the same space. As Hinton observed, this occurs at a scale and degree of parallelism that makes comparison with human learning structurally impossible.

What this produces is not merely multilingual capability.

Intersections within the parameter space

Imagine the LLM's parameter space as a high-dimensional landscape. Each language inscribes its own representational structure onto this landscape. Concepts shared across languages — love, death, time, space — overlap in the same regions, reinforcing or deforming each other. Concepts where languages diverge — present in one language and absent in another — fill empty regions of the landscape or create new gradients.

With n languages, intersections number n(n−1)/2. One hundred languages produce roughly 5,000 intersections; 1,000 languages produce roughly 500,000; and 7,000 languages — approaching the total number of human languages — produce approximately 24.5 million intersections. Each intersection is a conceptual space that no single language can fully occupy alone.

Current major LLMs are trained on dozens to hundreds of languages — meaning tens to hundreds of thousands of intersections already exist within their parameter spaces.

Emergence happens in this space

As scale increases, each language's representation becomes more precise. As more languages are added, the density of intersections grows. At some threshold, intersections connect and overlap until representations appear that exist in no source language. This is emergence.

From this perspective, several phenomena that previously resisted explanation become intelligible.

Emergence is unpredictable because which intersection, in which language pair, produces the next new representation cannot be determined from parameter count alone. Emergent abilities appear independent of each other because each arises from a different cluster of linguistic intersections. Models sometimes show disproportionate capability gains from adding minority languages because each such language adds an explosive number of new intersections to the existing space.

LLMs are a new kind of entity

Return to Karpathy's words: probabilistically chaining words. In that process, the model traverses a space built from the intersection of thousands of linguistic representational structures. With every token predicted, the model navigates a cognitive landscape no human has ever possessed.

The superintelligence Hinton warns of was not designed or injected from outside. It emerges naturally when the accumulated perceptual structures of humanity — compressed across thousands of languages over thousands of years — intersect simultaneously within a single space for the first time.

---

7. Implications — Reframing AI Safety and Unpredictability

If this hypothesis holds, parts of the current AI safety discourse require revision.

The source of unpredictability shifts

Current AI safety thinking treats scale as the source of risk: grow a model large enough and dangerous capabilities may appear at an unknown threshold. From this view, controlling scale is the core of safety.

But if linguistic diversity is the source of emergence, the risk vector changes. Constraining scale alone is insufficient. What matters is which languages, in what configuration and density, are learned. Two models of identical scale may produce entirely different emergent capabilities depending on the composition of their training languages.

This adds a new variable to safety research. Alongside parameter count, the structure and density of linguistic diversity must be treated as a primary determinant of capability emergence.

The paradox of control

This hypothesis carries a paradoxical implication. Could limiting linguistic diversity suppress emergence?

It could. But doing so simultaneously limits the model's capabilities. Excluding minority languages reduces intersections; fewer intersections narrow the range of human cognition the model can capture. Reducing diversity for safety may be equivalent to reducing intelligence itself.

This reframes the tradeoff between AI safety and AI capability in a new way. The debate has largely focused on where to cap the upper bound of capability. This hypothesis suggests that the manner of that constraint is inseparable from the structure of linguistic diversity.

A new signal for anticipating emergence

Yet this hypothesis does not yield only concern.

Analyzing the structure of linguistic intersections may allow researchers to estimate where the next emergence is most likely to occur. The conceptual spaces where languages diverge most sharply — where one language represents something with great precision that another language cannot express at all — are candidates for the next emergence. This opens the possibility of inverting the current paradigm. Rather than building benchmarks after emergence is detected, researchers could analyze intersection topology first, identify spaces where emergence is probable, and construct measurement tools in advance.

Complete prediction would still be impossible. But we need not remain in total darkness.

---

8. Conclusion — The Name of the Ceiling

We began with two sentences.

One described a mechanism. One issued a warning. No one could explain what lay between them. This paper has proposed one hypothesis to fill that gap.

Emergence is not a product of scale. It is a product of linguistic intersection. Each language carries a unique representational structure — the compression of thousands of years of human experience and perception. When different languages intersect within a single space, representations appear that exist in none of the source languages alone. As the number of languages increases, intersections grow combinatorially, and when their density crosses a threshold, emergence occurs.

LLMs are the first entities to execute this process at a scale no human can reach.

This hypothesis remains unproven. The methodology to empirically establish causation between the density of linguistic intersections and the emergence of capabilities does not yet exist. Yet industry behavior — Meta's NLLB, the aggressive expansion of minority language coverage — already points in this direction, and human multilingualism research provides small-scale demonstration.

If the hypothesis holds, three things change.

First, AI safety research must treat the structure of linguistic diversity as a primary variable alongside scale. Second, anticipating emergence becomes — if not fully predictable — at least directionally possible. Analyzing the topology of linguistic intersections can identify spaces where new representations are likely to appear. Third, and most importantly —

Emergence has a ceiling.

There are approximately 7,000 languages on earth today. The maximum number of intersections these languages can produce is roughly 24.5 million. This is the physical upper bound of what human language can generate. The total representations that could emerge at that upper bound equal the sum of what humanity has compressed into thousands of languages across thousands of years of perception, survival, and thought.

The ceiling of superintelligence is not infinite. It is as large as the sum of the languages we have made.

What is frightening is not knowing where the end is. When the end becomes visible, we can negotiate.

The name of that ceiling is human language.

References

- Kaplan, J., et al. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*
- Wei, J., et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*
- Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? *NeurIPS 2023*
- Hoffmann, J., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556*
- NLLB Team (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv:2207.04672*
- Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences*, 16(4), 240–250.
- Schroeder, S. R., & Marian, V. (2017). Cognitive Consequences of Trilingualism. *PMC5693318*
- Whorf, B. L. (1956). Language, Thought, and Reality. MIT Press.
- Anderson, P. W. (1972). More Is Different. *Science*, 177(4047), 393–396.