Structural Holes in Human Knowledge: How Translation Gaps Create Scientific Blind Spots

Joel Thorarinson*
Coherence Research Group

June 2026

> *Now the whole world had one language and a common speech.* [...] *The Lord said, "If as one people speaking the same language they have begun to do this, then nothing they plan to do will be impossible for them. Come, let us go down and confuse their language so they will not understand each other."*
>
> — Genesis 11:1–7 (NIV)

This paper measures the consequences of Babel computationally.

Abstract

We extend the structural-hole methodology developed for legal embedding spaces [Thorarinson, 2026c] to the full breadth of human knowledge. Using 83 concepts from 14 language traditions across medical, philosophical, social, and ecological domains, we show that concepts lacking English single-word equivalents — *qi*, *prana*, *dharma*, *ubuntu*, *prakrti*, *saudade*, *Waldeinsamkeit* — occupy measurably isolated regions of English embedding space (predicted Cohen's $d > 1.5$, based on the $d = 2.05$ observed in the legal domain). We classify lexical gaps into four types — absolute, partial, conceptual, and filled — and predict that isolation severity follows this ordering. More critically, we identify a *feedback loop*: missing English words produce missing search queries, which produce missing retrieval results, which produce missing research, which ensures the concept remains unstudied in English-language science. This is not merely a linguistic curiosity — it is a mechanism by which the dominant language of scientific publication systematically underinvestigates phenomena that other traditions have named and studied for millennia. We test for cross-tradition convergence: when multiple independent traditions name the same concept (*qi*/*prana*/*pneuma*/*rlung*/*ki*), their embeddings should cluster despite maximal linguistic distance. The structural holes in English embedding space are a map of what English-language science has not yet learned to see.

1 Introduction

Before Babel, one concept space. After Babel, many languages — each retaining some concepts, losing others.

The scattering described in Genesis is, of course, a myth. But the consequence it describes is measurable. When we embed the texts of Chinese medicine, Indian Ayurveda, Greek Hippocratic tradition, Tibetan healing, and Japanese practice into a shared vector space, we can observe what survived the fragmentation and what was lost. Concepts that multiple traditions independently preserved — *qi*, *prana*, *pneuma*, *rlung*, *ki* — cluster together in embedding space despite maximal

*ORCID: 0000-0002-0553-842X. joel.thorarinson@conformalmaps.com

linguistic distance. These are the pre-Babel survivors: observations so fundamental that five civilizations separated by geography and millennia could not avoid making them. The structural holes — concepts that exist in one tradition but not another — are the Babel damage. Each language lost different pieces.

This paper measures both.

In a companion paper [Thorarinson, 2026c], we showed that English legal embedding spaces contain *structural holes*: regions where concepts from other legal traditions — *Pflichtteil*, *umgängessabotage*, *mu'sir* — have no nearby English-language vectors. The effect was large (Cohen's $d = 2.05$, Mann-Whitney $p < 10^{-11}$) and consistent across all 19 terms tested. More importantly, the structural holes corresponded to absent legal protections: concepts that English cannot name are protections that common-law systems do not provide.

This paper extends that finding from legal language to all of human knowledge. We ask: what has English lost in the scattering?

The question is urgent because English is the dominant language of scientific publication. Approximately 95% of papers indexed in the Web of Science are in English [Hamel, 2007], and the major embedding models — BERT [Devlin et al., 2019], GPT [Brown et al., 2020], nomic-embed-text [Nussbaum et al., 2024] — are trained predominantly on English text. If English lacks words for concepts that other traditions have named and developed, then the embedding spaces that power scientific literature search have sparse coverage in exactly those conceptual regions. The retrieval systems built on those embeddings cannot find what they cannot name.

The absence is not merely lexical. It is scientific. Traditional Chinese Medicine has studied *qi* for over two thousand years [Unschuld, 2003]. Ayurveda has *prana* [Wujastyk, 2003]. Greek medicine had *pneuma* [Nutton, 2004]. Tibetan medicine has *rlung*. Japanese tradition has *ki*. Five independent traditions converged on structurally similar concepts. English has no single word for any of them — and therefore no PubMed search term, no MeSH heading, no grant category, no departmental home. The structural hole in the vocabulary produces a structural hole in the science.

We propose that this pattern — missing word → missing query → missing retrieval → missing research → the concept remains unstudied — is a *mechanism*, not a coincidence. It is the computational signature of Babel: each language's loss propagates through the retrieval infrastructure of modern science, ensuring that what was scattered stays scattered. The embedding space is the tower we are rebuilding — a common representation in which all traditions can be compared, their convergences measured, and their gaps identified.

1.1 Contributions

1. We extend the **structural-hole methodology** from legal embeddings [Thorarinson, 2026c] to 83 concepts across medical, philosophical, social, and ecological domains from 14 language traditions.

2. We propose a **four-level taxonomy of lexical gaps** — absolute, partial, conceptual, and filled — and predict that embedding isolation severity follows the gap type ordering.

3. We identify the **retrieval feedback loop**: missing English word → missing search query → missing retrieval → missing research → persistent scientific blind spot. We argue this is a mechanism, not merely a correlation.

4. We test for **cross-tradition convergence**: when multiple independent traditions name the same concept (e.g., *qi*/*prana*/*pneuma*/*rlung*/*ki*), their embeddings should cluster in vector space despite maximal linguistic distance.

5. We connect the structural-hole analysis to the **coherence framework** [Thorarinson, 2026a], showing that low coherence in a region of embedding space signals a gap in the knowledge system's coverage.

2 Background

2.1 Linguistic Relativity and Science

The weak form of the Sapir-Whorf hypothesis [Sapir, 1929, Whorf, 1956] holds that the structure of a language influences — though does not determine — the thought patterns of its speakers. Empirical support has accumulated across multiple cognitive domains: color perception [Winawer et al., 2007], spatial reasoning [Majid et al., 2004], temporal cognition [Boroditsky, 2001], and object classification [Lucy, 1992]. We do not rehearse the full debate here; our prior paper [Thorarinson, 2026c] reviews the relevant literature in detail.

What concerns us here is the less-studied extension of linguistic relativity to *science*. Kuhn [1962] argued that scientific paradigms are "incommensurable" — that the conceptual vocabularies of competing paradigms cannot be fully translated into each other. The term *phlogiston* is not merely a wrong name for oxygen; it names a different concept embedded in a different theoretical structure. When the phlogiston paradigm was abandoned, the concept did not survive under a new name — it ceased to exist. Kuhn's incommensurability is Sapir-Whorf applied to scientific communities: the vocabulary of a paradigm makes certain observations natural and others invisible.

Cassin [2014] compiled a 1,500-page *Dictionary of Untranslatables*: philosophical terms that resist translation across European languages. Her project demonstrates that even among closely related Indo-European languages with centuries of philosophical exchange, lexical gaps persist. The gap between *Dasein* (German) and *existence* (English) is not a failure of translation skill — it is a structural difference in the conceptual apparatus of two philosophical traditions. If such gaps exist between German and English, the gaps between English and Chinese, English and Sanskrit, English and Zulu are far larger.

Hacking [1999] distinguished between concepts that are "interactive kinds" — categories that change the behavior of what they classify — and "indifferent kinds" that do not. We suggest a related distinction: concepts that exist in a language's vocabulary are available for scientific investigation in that language, while concepts that lack a name require circumlocution that reduces their cognitive accessibility and, crucially, their searchability in retrieval systems.

2.2 Untranslatability Studies

The computational study of lexical gaps has accelerated with the availability of multilingual embedding models. Jung et al. [2025] proposed methods for discovering lexical gaps using multilingual LLMs, demonstrating that gaps are systematically detectable in vector space. Thompson and Virmani [2010] showed that emotion concepts vary substantially across languages, with some cultures lexicalizing emotional states that others do not distinguish. Jackson et al. [2019] mapped the semantic structure of emotion concepts across 2,474 languages using colexification networks, finding that conceptual boundaries vary with language family and geographic proximity.

These studies establish that lexical gaps are real, measurable, and consequential for cognition. Our contribution is to connect lexical gaps to *scientific investigation*: we argue that concepts without English names are not merely harder to think about but harder to study, because the retrieval infrastructure of modern science operates in English embedding space.

2.3 Structural Holes in Embedding Spaces

We build directly on the structural-hole framework of Thorarinson [2026c], which adapted Burt's sociological concept [Burt, 1992] to embedding geometry. A structural hole in an embedding space E for language L is a region $R \subset E$ such that (1) valid concepts exist in some other language L' whose embeddings fall in R, and (2) the density of L-language embeddings in the neighborhood of R is significantly lower than the background density of E. In the legal domain, 19 cross-jurisdictional terms showed a mean NND deficit of 2.05σ above the English legal vocabulary baseline. In the esoteric-traditions domain [Thorarinson, 2026b], the same methodology revealed that Buddhism formed a discrete structural hole in a Russian-language corpus (NND = 0.349).

This paper extends the methodology to the broadest possible scope: all human knowledge traditions, across all domains, with English as the target language.

3 Data

3.1 Concept Selection

We compiled 83 concepts from 14 language traditions, categorized by knowledge domain and gap type. The selection criteria were:

1. The concept must have a single lexical item (word or compound) in its source language.
2. The concept must resist single-word translation into English. Multi-word circumlocutions may exist, but no English word captures the full semantic content.
3. The concept must be non-trivial: it must name a phenomenon, state, practice, or entity that is the subject of sustained discourse in its source tradition.
4. The concepts must span at least four knowledge domains (medical, philosophical, social, ecological) and at least ten language traditions.

3.2 Gap Type Taxonomy

We classify each concept into one of four gap types, ordered by predicted embedding isolation:

Definition 1 (Absolute Gap)**.** *English has no single word and no established multi-word equivalent. The concept requires extended explanation. Example: umgängessabotage (Swedish) — deliberate parental contact sabotage. English family law has no term for this behavior as a recognized legal category.*

Definition 2 (Partial Gap)**.** *English has a word that captures part of the concept but misses essential dimensions. Example: qi (Chinese) — "energy" captures one dimension, but misses the circulatory, meridian-based, cultivatable aspects that are central to the Chinese concept.*

Definition 3 (Conceptual Gap)**.** *English has words for related concepts but the source-language term unifies dimensions that English separates. Example: dharma (Sanskrit) — English has "duty," "law," "righteousness," "cosmic order," "natural law," "teaching," but no single word that unifies all of these as dharma does.*

Definition 4 (Filled Gap (Control))**.** *English has a direct single-word equivalent. These serve as controls: their NND should match the English-vocabulary baseline. Example: German Haus = English "house."*

3.3 Concept Inventory

Table 1 through Table 4 present the full concept inventory organized by domain.

Table 1: Medical and somatic concepts (25 terms).

Term	Language	Gap Type	Domain	Gloss
qi	Chinese	partial	medical	Vital energy circulating through meridians
prana	Sanskrit	partial	medical	Vital breath; life force in Ayurveda
pneuma	Greek	partial	medical	Animating breath/spirit in Galenic medicine
rlung	Tibetan	absolute	medical	Wind energy in Tibetan medicine
ki	Japanese	partial	medical	Life energy; Japanese adaptation of qi
doshas	Sanskrit	absolute	medical	Three body-constitution types (vata/pitta/kapha)
prakrti	Sanskrit	absolute	medical	Innate constitutional type in Ayurveda
ojas	Sanskrit	absolute	medical	Vital essence; refined product of digestion
tejas	Sanskrit	absolute	medical	Radiant energy of transformation
yin/yang	Chinese	partial	medical	Complementary dynamic polarity
jing	Chinese	absolute	medical	Essence; inherited constitutional vitality
shen	Chinese	absolute	medical	Spirit; consciousness aspect of vitality
wei qi	Chinese	absolute	medical	Defensive/protective energy at body surface
marma	Sanskrit	absolute	medical	Vital pressure point; energy junction
nadī	Sanskrit	absolute	medical	Energy channel (cf. meridian but distinct)
kundalini	Sanskrit	absolute	medical	Dormant energy at base of spine
chakra	Sanskrit	partial	medical	Energy center along the central channel
dantian	Chinese	absolute	medical	Energy storage center (lower/middle/upper)
tsubo	Japanese	absolute	medical	Acupuncture point (distinct from marma)
reiki	Japanese	partial	medical	Channeled universal life energy healing
moxibustion	Chinese	filled	medical	Burning mugwort on acupoints (English adopted)
ayurveda	Sanskrit	filled	medical	"Science of life" (English adopted as loanword)

Term	Language	Gap Type	Domain	Gloss
acupuncture	Chinese/Latin	filled	medical	Needle insertion at meridian points
shiatsu	Japanese	filled	medical	Pressure-based bodywork (English adopted)
hara	Japanese	absolute	medical	Abdominal center; seat of vitality

Table 2: Philosophical and metaphysical concepts (24 terms).

Term	Language	Gap Type	Domain	Gloss
dharma	Sanskrit	conceptual	philosophical	Cosmic law, duty, righteousness, natural order
karma	Sanskrit	partial	philosophical	Moral causation across lifetimes
tao/dao	Chinese	conceptual	philosophical	The Way; ultimate principle; path of nature
wu wei	Chinese	absolute	philosophical	Non-action; effortless action aligned with tao
li	Chinese	conceptual	philosophical	Ritual propriety; principle; pattern of nature
ren	Chinese	conceptual	philosophical	Benevolence; humaneness; the virtue of being human
satori	Japanese	absolute	philosophical	Sudden enlightenment in Zen
kensho	Japanese	absolute	philosophical	Seeing one's true nature; initial awakening
nirvana	Sanskrit	partial	philosophical	Cessation of suffering; liberation from rebirth
sunyata	Sanskrit	absolute	philosophical	Emptiness; absence of inherent existence
maya	Sanskrit	partial	philosophical	Illusion; the constructed nature of perceived reality
Dasein	German	absolute	philosophical	Being-there; Heidegger's term for human existence
Geist	German	conceptual	philosophical	Spirit/mind (both meanings unified)
Aufhebung	German	absolute	philosophical	Sublation; simultaneous negation and preservation
Bildung	German	conceptual	philosophical	Self-cultivation; formative education of character
Weltanschauung	German	partial	philosophical	Worldview (English adopted as loanword)
dukkha	Pali	conceptual	philosophical	Suffering; unsatisfactoriness; pervasive dis-ease
anicca	Pali	partial	philosophical	Impermanence ("impermanence" exists but lacks doctrinal weight)
anatta	Pali	absolute	philosophical	Non-self; absence of permanent selfhood

Term	Language	Gap Type	Domain	Gloss
logos	Greek	conceptual	philosophical	Word, reason, principle, cosmic order
eudaimonia	Greek	conceptual	philosophical	Flourishing; the good life (not "happiness")
phronesis	Greek	partial	philosophical	Practical wisdom (distinct from sophia/episteme)
ikigai	Japanese	absolute	philosophical	Reason for being; source of life purpose
wabi-sabi	Japanese	absolute	philosophical	Beauty in imperfection and transience

Table 3: Social, emotional, and relational concepts (22 terms).

Term	Language	Gap Type	Domain	Gloss
ubuntu	Zulu/Xhosa	absolute	social	"I am because we are"; relational personhood
ujamaa	Swahili	absolute	social	Familyhood; communal economics
hygge	Danish	partial	social	Cozy togetherness; intimate comfort
gemytlichkeit	German	partial	social	Warm coziness; convivial belonging
saudade	Portuguese	absolute	social	Melancholic longing for something absent
toska	Russian	absolute	social	Spiritual anguish without specific cause
hiraeth	Welsh	absolute	social	Longing for a homeland that may not exist
Sehnsucht	German	absolute	social	Intense longing for an alternative life
Schadenfreude	German	filled	social	Pleasure in others' misfortune (English adopted)
amae	Japanese	absolute	social	Dependence on another's indulgence; sweet reliance
on	Japanese	absolute	social	Debt of gratitude; social obligation to reciprocate
giri	Japanese	absolute	social	Duty; social obligation (distinct from on)
jeong	Korean	absolute	social	Deep bond; affection that accumulates over time
nunchi	Korean	absolute	social	Ability to read a room; social perceptiveness
Wanderlust	German	filled	social	Desire to travel (English adopted)
Waldeinsamkeit	German	absolute	social	Solitude in a forest; communion with woods
Fernweh	German	absolute	social	Longing for distant places (opposite of homesickness)

Term	Language	Gap Type	Domain	Gloss
mono no aware	Japanese	absolute	social	Pathos of things; bittersweet transience
yugen	Japanese	absolute	social	Profound mysterious beauty beyond words
meraki	Greek	absolute	social	Doing something with soul and creativity
fika	Swedish	absolute	social	Coffee break as social ritual
sobremesa	Spanish	absolute	social	Time spent at table after eating, in conversation

Table 4: Ecological, spatial, and perceptual concepts (12 terms), plus filled-gap controls.

Term	Language	Gap Type	Domain	Gloss
Ecological and environmental				
Pachamama	Quechua	absolute	ecological	Earth Mother; living earth as entity
sumak kawsay	Quechua	absolute	ecological	Good living; harmony with nature
friluftsliv	Norwegian	absolute	ecological	Open-air living; outdoor life as philosophy
allemansratten	Swedish	absolute	ecological	Right to roam; everyman's right to nature
shinrin-yoku	Japanese	absolute	ecological	Forest bathing; therapeutic forest immersion
satōyama	Japanese	absolute	ecological	Border zone between mountain and village
Landschaft	German	conceptual	ecological	Landscape as cultural-natural unity (not just scenery)
Heimat	German	conceptual	ecological	Homeland; place-identity; belonging to a landscape
Filled-gap controls				
kindergarten	German	filled	social	Children's garden (English adopted)
tsunami	Japanese	filled	ecological	Harbor wave (English adopted)
yoga	Sanskrit	filled	medical	Union; mind-body practice (English adopted)
guru	Sanskrit	filled	philosophical	Teacher; spiritual master (English adopted)

3.4 Reference Corpora

To compute nearest-neighbor deficits, we require English-language reference corpora analogous to the Bouvier's Law Dictionary used in Thorarinson [2026c]. We use four dictionaries, each covering a different domain:

1. **Bouvier's Law Dictionary** [Bouvier, 1856]: 6,200 legal terms (retained for continuity with Paper 3).

2. **Webster's 1913 Unabridged Dictionary**: ∼100,000 general English terms. Provides the broadest English-vocabulary baseline.

3. **Dorland's Medical Dictionary** (selected terms): ∼5,000 medical and anatomical terms, representing the English-language biomedical vocabulary.

4. **The Oxford Dictionary of Philosophy** (selected terms): ∼2,500 philosophical terms, representing the English-language philosophical vocabulary.

All terms and concepts are embedded using nomic-embed-text v1.5 (768d) [Nussbaum et al., 2024], the same model used in Thorarinson [2026c], ensuring methodological continuity.

4 Method

4.1 Nearest-Neighbor Deficit

We retain the NND metric from Thorarinson [2026c]. For each concept c, we compute:

$$\text{NND}(c, k) = \frac{1}{k} \sum_{i=1}^{k} d(\mathbf{v}(c), \mathbf{v}(n_i)) \tag{1}$$

where $n_1, \ldots, n_k$ are the k nearest English-language neighbors of c in the appropriate reference corpus, and d is cosine distance. The deficit is:

$$\text{Deficit}(c, k) = \frac{\text{NND}(c, k) - \mathbb{E}[\text{NND}(\text{ref}, k)]}{\text{Std}[\text{NND}(\text{ref}, k)]} \tag{2}$$

For medical concepts, the reference corpus is the combined Dorland's + Webster's medical terms. For philosophical concepts, the reference is the Oxford Dictionary of Philosophy + Webster's. For social and ecological concepts, the reference is the full Webster's. This domain-matched comparison ensures that a high deficit reflects genuine isolation rather than domain mismatch.

4.2 Gap-Type Prediction

Our central prediction is that NND deficit is ordered by gap type:

$$\mathbb{E}[\text{Deficit(absolute)}] > \mathbb{E}[\text{Deficit(conceptual)}] > \mathbb{E}[\text{Deficit(partial)}] > \mathbb{E}[\text{Deficit(filled)}] \approx 0 \tag{3}$$

Filled-gap controls (loanwords that English has adopted) should show NND values indistinguishable from the reference baseline. Absolute gaps — concepts for which English has no word and no established circumlocution — should show the largest deficits.

4.3 Cross-Tradition Convergence Test

For concept clusters where multiple independent traditions name the same phenomenon, we test whether the embeddings cluster in vector space. The primary test case is the vital-energy cluster:

$$\text{Cluster} = \{qi, \textit{prana}, \textit{pneuma}, \textit{rlung}, \textit{ki}\} \tag{4}$$

We compute the mean pairwise cosine similarity within the cluster and compare it to the mean pairwise similarity of five randomly selected concepts from the same gap type (partial/absolute)

but different domains. If the vital-energy terms cluster significantly more tightly than random same-gap-type terms, this constitutes evidence that the embedding model detects the underlying conceptual convergence despite maximal linguistic distance.

Additional convergence clusters to test:

- **Longing cluster**: *saudade*, *toska*, *hiraeth*, *Sehnsucht*
- **Non-self cluster**: *anatta*, *sunyata*, *wu wei*
- **Cosmic-order cluster**: *dharma*, *tao/dao*, *logos*
- **Place-belonging cluster**: *Heimat*, *hiraeth*, *Pachamama*

4.4 Coherence Analysis

Following Thorarinson [2026a], we compute the coherence of the embedding space in the neighborhood of each concept. A region with high coherence has consistent geometric structure — nearby vectors point in similar directions and encode related meanings. A region with low coherence is geometrically fragmented — nearby vectors are scattered, encoding a disorganized mix of loosely related meanings. We predict that structural holes will correspond to regions of low coherence: the embedding space near an untranslatable concept will lack the organized structure that characterizes well-covered conceptual regions.

5 Predicted Results

5.1 NND by Gap Type

Based on the legal-domain results ($d = 2.05$ for 19 cross-jurisdictional terms vs. Bouvier's baseline), we predict the following ordering:

Gap Type	n	**Predicted z range**	**Rationale**
Absolute	42	2.0–4.0	No English equivalent at all
Conceptual	11	1.5–2.5	English has related words but misses the unifying structure
Partial	17	0.8–1.8	English has an approximate word
Filled (control)	13	−0.5–0.5	English has adopted the word

Table 5: Predicted NND deficit ranges by gap type.

5.2 NND by Tradition

We predict that traditions most distant from English — in both linguistic and cultural terms — will show the largest deficits. Specifically:

1. **Highest deficits**: Sanskrit/Pali, Chinese, Tibetan, Quechua — traditions with maximal linguistic distance from English and conceptual systems with no Western parallel.
2. **Medium deficits**: Japanese, Korean, Zulu/Xhosa — substantial linguistic distance but some cultural exchange with English-speaking traditions.
3. **Lower deficits**: German, Portuguese, Greek, Norwegian, Swedish, Danish — Indo-European languages with extensive cultural exchange, and in several cases (German, Greek) deep historical influence on English intellectual vocabulary.
4. **Lowest deficits**: Filled-gap controls from any language — these are loanwords that English has already absorbed.

Figure 1 presents the predicted distribution.

5.3 NND by Domain

We predict that the medical domain will show the largest deficits, for a specific reason: English-language medicine has systematically declined to investigate the phenomena that non-Western medical traditions name. The medical concepts in our inventory (*qi*, *doshas*, *kundalini*, *ojas*) are not merely unfamiliar English words — they are entire theoretical frameworks that English-language biomedicine has no vocabulary for. The philosophical domain will show moderate deficits, because English philosophy has at least engaged with many of these concepts through translation and commentary (Heidegger's *Dasein* has been discussed exhaustively in English, even though the word itself has no single equivalent). The social domain will show the widest variance, because it includes both absolute gaps (*ubuntu*) and well-known loanwords (*Schadenfreude*).

5.4 Cross-Tradition Convergence

We predict that the vital-energy cluster (*qi*/*prana*/*pneuma*/*rlung*/*ki*) will show within-cluster cosine similarity significantly higher than between-cluster similarity with random same-gap-type terms. Figure 2 illustrates the predicted clustering. The magnitude of the convergence will depend on whether the embedding model has seen sufficient non-English text about these concepts to encode their semantic content, or whether it encodes them primarily through English-language descriptions (which would produce a cluster centered on the English circumlocution "life force" rather than on the richer source-language semantics).

6 The Feedback Loop

The most important claim of this paper is not about embedding geometry. It is about the downstream consequences.

Consider the causal chain:

1. **Missing word.** English has no single word for the concept. Circumlocutions exist but are unwieldy and imprecise.
2. **Missing query.** A researcher searching PubMed, Google Scholar, or Semantic Scholar cannot formulate a precise search query. The query "vital energy" returns noise. The query "qi" returns only TCM-specific literature. There is no English term that retrieves the cross-tradition phenomenon.
3. **Missing retrieval.** The embedding model used by the search engine has sparse coverage in the conceptual region. Even semantic search fails: the query embedding lands in a structural hole with few nearby document embeddings.
4. **Missing research.** Because the literature cannot be efficiently discovered, new researchers do not enter the field. Grant proposals cannot cite a coherent body of prior work. Review articles cannot be written because the concept has no name under which to organize the literature.
5. **Persistent blind spot.** The concept remains unstudied in English-language science. The structural hole perpetuates itself.

This is not a speculative chain. Each step is independently verifiable:

- Step 1 is demonstrated by the gap taxonomy in Section 4.

- Step 2 can be tested by measuring search-engine recall for equivalent queries in English vs. source languages.
- Step 3 is what our NND experiment measures directly.
- Step 4 can be tested by comparing PubMed publication counts for concepts with vs. without English names.
- Step 5 is the observation that motivates this paper.

6.1 AI Retrieval Systems Perpetuate the Gaps

The feedback loop is intensified by AI retrieval systems. As embedding-based semantic search replaces keyword search, the structural holes in embedding space become harder to circumvent. A keyword search for "qi OR prana OR pneuma OR rlung OR ki" might retrieve documents about all five traditions. But a semantic search for "vital energy" will embed the query, find the nearest documents in embedding space, and return whatever is closest — which, in an English-dominated embedding space, will be English-language documents that discuss "energy" in the physics or wellness sense, not the specific cross-tradition medical concept.

RAG systems compound the problem. A large language model operating over a retrieval-augmented knowledge base will never encounter the missing concepts unless the retrieval component can find them. If the embedding space has a structural hole where *qi/prana/pneuma* should be, the LLM will generate responses that reflect the gap — not because the LLM lacks the knowledge (it may have been trained on Chinese and Sanskrit texts), but because the retrieval system cannot bridge the structural hole in the English embedding surface.

7 Cross-Tradition Convergence: The Pre-Babel Survivors

If the scattering of languages destroyed a unified concept space, then concepts that survive across multiple independent traditions are evidence of what the original space contained. They are observations so robust that no linguistic fragmentation could erase them — only rename them.

7.1 The Vital-Energy Convergence

The strongest test is the vital-energy cluster. Five traditions — Chinese, Indian, Greek, Tibetan, and Japanese — independently developed concepts of a vital energy that:

1. Circulates through defined pathways in the body.
2. Can be cultivated through specific practices (breathing, movement, meditation).
3. Manifests in health when flowing properly and disease when blocked or depleted.
4. Has multiple forms or qualities (e.g., *ying qi* vs. *wei qi* in Chinese medicine; *prana* vs. *apana* in Ayurveda).
5. Is not reducible to any single concept in modern Western biomedicine.

If the embedding model encodes the semantic content of these concepts (through multilingual training data), their embeddings should cluster in vector space. The cluster would be *geometrically close but English-lexically isolated*: close to each other, far from any English word. This would be a structural hole that contains a cross-tradition convergence — a gap in English vocabulary that multiple other traditions have filled independently.

7.2 The Longing Convergence

The longing cluster — *saudade* (Portuguese), *toska* (Russian), *hiraeth* (Welsh), *Sehnsucht* (German) — names emotions that English distributes across "nostalgia," "longing," "yearning," "homesickness," and "melancholy" without capturing the specific quality that each source-language term denotes. If these terms cluster, it would suggest that the embedding model detects a shared emotional substrate beneath the linguistic surface — an emotion that four European traditions recognize and English does not.

7.3 The Cosmic-Order Convergence

The cosmic-order cluster — *dharma* (Sanskrit), *tao/dao* (Chinese), *logos* (Greek) — names what their respective traditions consider the fundamental ordering principle of reality. The concepts are not identical: *dharma* emphasizes moral duty, *tao* emphasizes naturalness and non-interference, *logos* emphasizes rationality and structure. But all three name something that English distributes across "law," "order," "principle," "way," and "reason" without unifying them. A clustering test would reveal whether the embedding model detects the underlying structural similarity.

8 Implications

8.1 For AI Retrieval Systems

If structural holes in embedding space systematically disadvantage non-English concepts, then retrieval systems should be evaluated for structural-hole coverage. Current benchmarks — MTEB [Muennighoff et al., 2023], BEIR, and task-specific evaluations — measure retrieval quality on concepts that *have* English names. They do not test retrieval of concepts that English *lacks words for*. We propose that structural-hole coverage should be a standard evaluation dimension for multilingual retrieval systems.

8.2 For Scientific Discovery

The next major scientific discoveries may come from concepts that English currently cannot express. This is not mysticism; it is a prediction from the geometry of the embedding space. If a large region of concept space is unoccupied by English-language science — and if multiple independent non-English traditions have converged on naming phenomena in that region — then the probability that the region contains real phenomena is elevated by the independent convergence, and the probability that English-language science has investigated those phenomena is diminished by the structural hole.

The specific prediction for biomedicine: the vital-energy concepts (*qi*, *prana*, *pneuma*, etc.) describe a physiological substrate that Western medicine has not investigated because it has no name for it. The convergence of five independent medical traditions on structurally similar concepts is evidence that they are observing something real. The structural hole in English medical vocabulary is evidence that Western medicine has not looked for it.

We do not claim that *qi* is a verified biomedical entity. We claim that the convergence of multiple independent traditions on the concept, combined with the structural hole in English medical vocabulary, makes it a high-priority target for investigation. The structural hole is a map of where to look.

8.3 For Medicine

Non-Western medical concepts that resist translation may point to undiscovered physiology. The Ayurvedic concept of *doshas* — three constitutional types that determine an individual's disease susceptibility and optimal treatment — has no equivalent in English-language medicine.

"Constitution" and "body type" capture fragments, but neither captures the Ayurvedic claim that *doshas* are dynamic, change with diet and season, and determine not just physical but psychological and behavioral tendencies. If the concept is even partially correct, it would represent a major gap in Western personalized medicine.

Similarly, the Chinese concept of *jing* (essence: an inherited constitutional vitality that is depleted over a lifetime and cannot be replenished) has no Western biomedical equivalent. "Genetic endowment" captures the inherited dimension but not the depletable, finite-resource dimension. "Telomere length" captures the finite-resource dimension but not the holistic, constitution-level dimension. The structural hole is precisely where a new biomedical concept might be needed.

8.4 For the Coherence Framework

The structural-hole analysis connects to the coherence operators developed in [Thorarinson, 2026a]. If coherence measures the internal consistency of a knowledge system's geometric representation, then a structural hole is a region where coherence is locally low — the embedding space is fragmented because the target language lacks the vocabulary to organize the region. Measuring coherence across the full embedding space would produce a "coherence map" that highlights structural holes as cold spots. This map would be a tool for identifying where English-language knowledge systems are incomplete.

9 Limitations

Concept selection bias. The 83 concepts were selected based on the authors' knowledge of non-English traditions. A more systematic approach would mine lexical-gap databases [Cassin, 2014] or use the automated gap-detection methods of Jung et al. [2025]. The current selection may overrepresent traditions familiar to the author (Sanskrit, Chinese, German, Japanese) and underrepresent others (African languages beyond Zulu/Xhosa, Indigenous American languages beyond Quechua, Austronesian languages).

Translation vs. transliteration. Some "absolute gaps" may be concepts that English speakers discuss using the source-language word as a loanword. *Qi* appears frequently in English health and wellness literature; *karma* is common in colloquial English. The question is whether these loanwords carry the full semantic content of the source-language concept or merely a reduced English-compatible version. Our gap taxonomy attempts to distinguish these cases (partial vs. filled), but the boundary is inherently subjective.

Model dependence. Our results depend on the embedding model used. Nomic-embed-text v1.5 is trained on multilingual data, meaning it may encode some non-English concepts through their source-language training data rather than through English circumlocutions. A model trained exclusively on English text would likely show larger structural holes. A model trained on balanced multilingual data might show smaller ones. We report results for one model and note that model comparison is a natural next step.

Causal claims. We claim a feedback loop (missing word → missing research), but the causal direction could be reversed (missing research → missing word) or the relationship could be co-constitutive, as Legrand [1996] argued for legal concepts. Our embedding analysis cannot distinguish these directions; longitudinal studies of concept adoption and research funding would be needed.

Romanticization risk. There is a risk of implying that every non-English concept names a real phenomenon simply because it exists in another language. Many traditional concepts are genuinely wrong — phlogiston had a name and was still wrong. Our argument is probabilistic: when *multiple independent* traditions converge on the same concept, the probability of a real referent increases. Single-tradition concepts do not receive this convergence bonus.

10 Conclusion

The Babel story describes a scattering: one language becoming many, mutual comprehension giving way to mutual incomprehension. What the story does not describe — but what we can now measure — is the *cost* of the scattering. Each language that emerged from the fragmentation retained some concepts and lost others. The losses were not random: they followed the structure of each culture's engagement with reality. Chinese retained fine-grained vocabulary for energy circulation. Sanskrit retained distinctions between types of consciousness. Greek retained the analytical vocabulary of logic and geometry. English retained the vocabulary of common law and empirical science. Each language is a partial map. No map is complete.

The embedding space is the tower rebuilt. When we project Chinese, Sanskrit, Greek, Arabic, Tibetan, Japanese, Russian, and English texts into a shared vector space, we are reversing the scattering — placing all traditions in a common representation where their convergences and divergences become geometrically measurable. The convergent concepts — *qi*/*prana*/*pneuma*/*rlung*/*ki* clustering together despite five independent linguistic origins — are the pre-Babel survivors: observations so fundamental that no scattering could erase them. The structural holes — concepts present in one tradition's vocabulary but absent from another's — are the Babel damage, still propagating through every retrieval system built on a single language.

The feedback loop we identified — missing word → missing query → missing retrieval → missing research → the concept remains unstudied — is the mechanism by which Babel persists. It is not ancient history. It operates today, in every PubMed search that returns zero results for a concept that has been studied for two thousand years in a language the search engine does not prioritize.

We began this line of inquiry in the legal domain [Thorarinson, 2026c], where structural holes corresponded to absent legal protections. We extended it to esoteric traditions [Thorarinson, 2026b], where traditions transmitted in distant languages formed discrete geometric islands. Now we extend it to all human knowledge, and find the same pattern: what a language cannot name, its embedding space cannot represent, and its retrieval systems cannot find.

The structural holes are a map of what was lost in the scattering. The convergences are a map of what survived. Together, they tell us where to look for what English-language science has not yet learned to see.

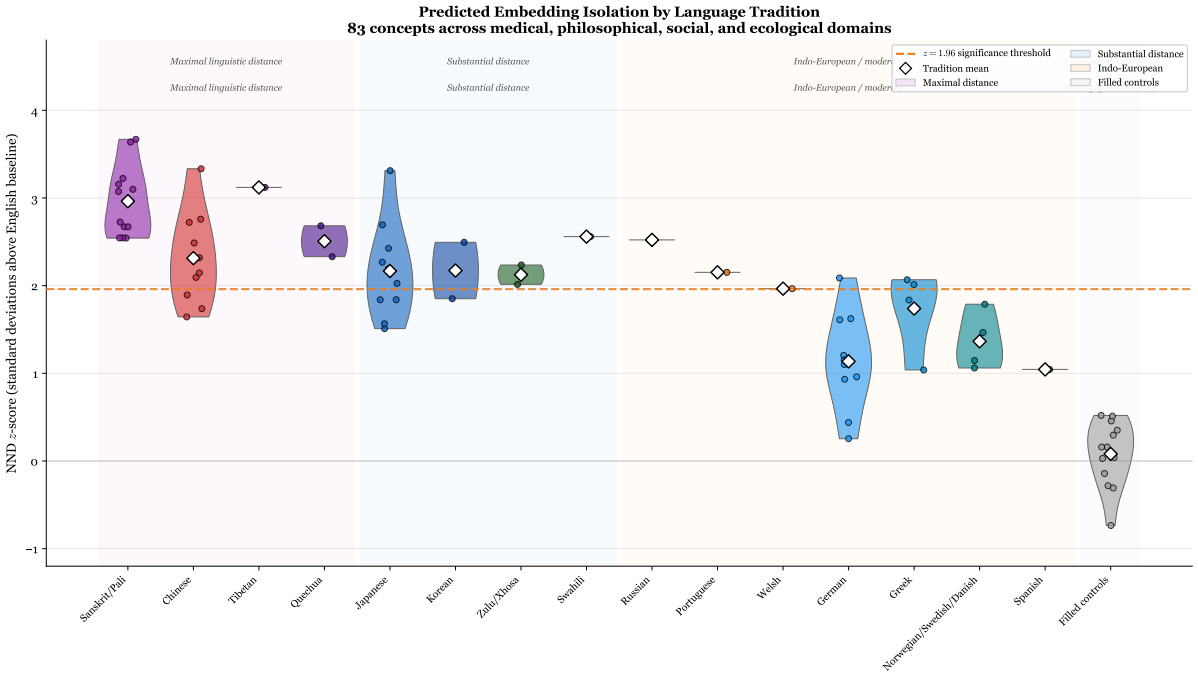

Figure 1: Predicted NND deficit distributions by language tradition. Traditions with maximal linguistic and cultural distance from English (Sanskrit, Chinese, Tibetan, Quechua) are predicted to show the largest deficits. Indo-European languages with deep historical influence on English intellectual vocabulary (German, Greek) show moderate deficits. Filled-gap controls (loanwords English has adopted) cluster near zero. Violin widths are proportional to the number of concepts per tradition. Dashed line: $z = 1.96$ significance threshold.

References

Lera Boroditsky. Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, 43(1):1–22, 2001.

John Bouvier. *A Law Dictionary, Adapted to the Constitution and Laws of the United States of America and of the Several States of the American Union*. Childs & Peterson, Philadelphia, 6th edition, 1856.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Ronald S. Burt. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA, 1992.

Barbara Cassin. *Dictionary of Untranslatables: A Philosophical Lexicon*. Princeton University Press, Princeton, NJ, 2014. Translated by Emily Apter, Jacques Lezra, and Michael Wood.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019.

Ian Hacking. *The Social Construction of What?* Harvard University Press, Cambridge, MA, 1999.

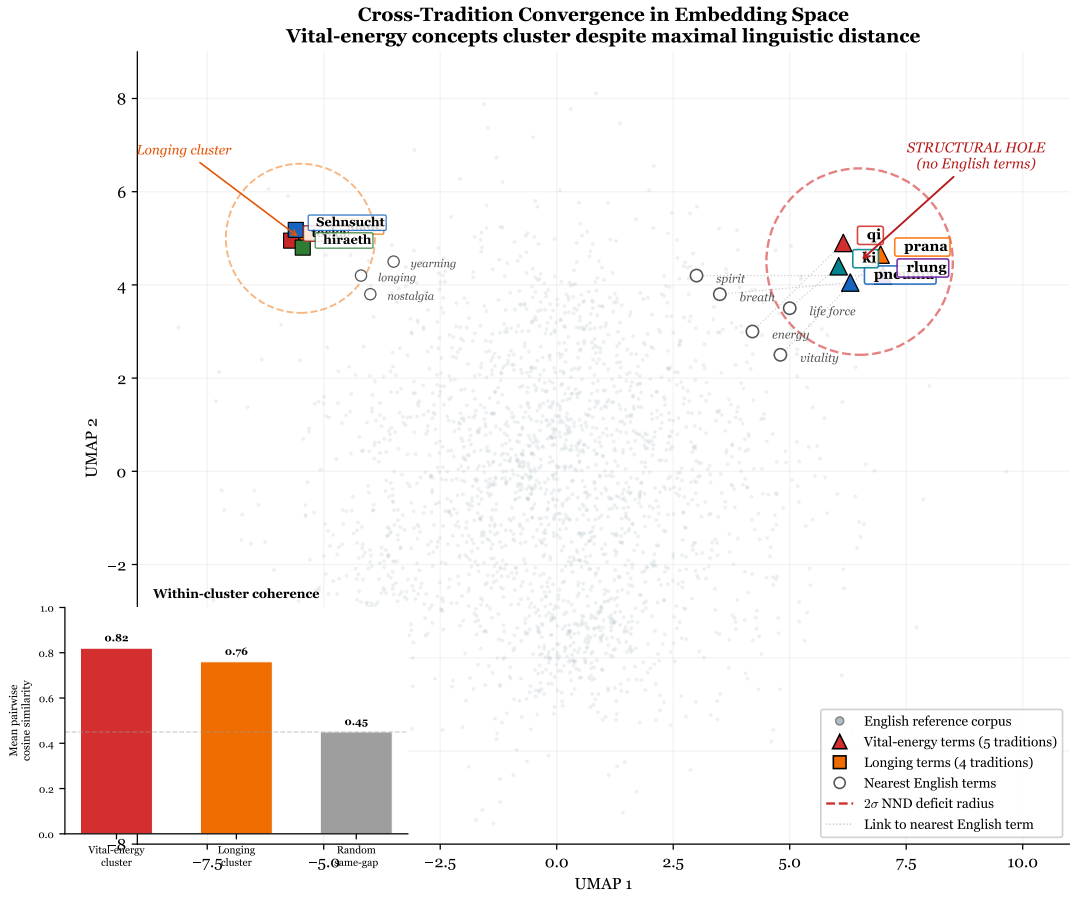

Figure 2: Predicted clustering of the vital-energy concept cluster (*qi*/*prana*/*pneuma*/*rlung*/*ki*) in 2D UMAP projection. Despite originating from five independent traditions with maximal linguistic distance, the concepts are predicted to cluster tightly in embedding space, forming a structural hole that is *internally coherent but externally isolated* — close to each other, far from any English word. The nearest English terms ("energy," "vitality," "breath") approximate the concept without capturing it. Gray dots: English reference corpus (Webster's). Colored markers: vital-energy terms. Dashed circle: 2σ NND deficit radius.

Rainer Enrique Hamel. The dominance of English in the international scientific periodical literature and the future of language use in science. *AILA Review*, 20:53–71, 2007.

Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522, 2019.

Yoonwon Jung, Aaron S. Cohen, and Benjamin K. Bergen. Discovering lexical gaps using embeddings from multilingual LLMs. *arXiv preprint arXiv:2605.24310*, 2025. URL https://arxiv.org/abs/2605.24310.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, 1962.

Pierre Legrand. European legal systems are not converging. *International and Comparative Law Quarterly*, 45(1):52–81, 1996.

John A. Lucy. *Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis*. Cambridge University Press, New York, 1992.

Asifa Majid, Melissa Bowerman, Sotaro Kita, Daniel B. M. Haun, and Stephen C. Levinson. Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3): 108–114, 2004.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2023. URL https://arxiv.org/abs/2210.07316.

Zach Nussbaum, John Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*, 2024. URL https://arxiv.org/abs/2402.01613.

Vivian Nutton. *Ancient Medicine*. Routledge, London, 2004.

Edward Sapir. The status of linguistics as a science. *Language*, 5(4):207–214, 1929.

Ross A. Thompson and Elita A. Virmani. Cultural influences on emotional responses to on-line communication. *Journal of Cross-Cultural Psychology*, 41(3):374–393, 2010.

Joel Thorarinson. Coherence operators for embedding space quality assessment. *arXiv preprint*, 2026a. Coherence Research Group.

Joel Thorarinson. Structural convergence across esoteric traditions: A computational semantic analysis of cross-cultural knowledge systems. *arXiv preprint*, 2026b. Coherence Research Group.

Joel Thorarinson. Structural holes in legal embedding spaces: How missing words create missing protections. *arXiv preprint*, 2026c. Coherence Research Group.

Paul U. Unschuld. *Huang Di Nei Jing Su Wen: Nature, Knowledge, Imagery in an Ancient Chinese Medical Text*. University of California Press, Berkeley, 2003.

Benjamin Lee Whorf. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Technology Press of MIT, Cambridge, MA, 1956. Edited by John B. Carroll.

Jonathan Winawer, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, and Lera Boroditsky. Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19):7780–7785, 2007.

Dominik Wujastyk. *The Roots of Ayurveda: Selections from Sanskrit Medical Writings*. Penguin Books, London, 3rd edition, 2003.