Frontiers of Architectural Research 15 (2026) 409–423

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.keaipublishing.com/foar

RESEARCH ARTICLE

Navigating CLIPedia: Architectonic instruments for querying and questing a latent encyclopedia

Agostino Nickl

Institute of Technology in Architecture, Department of Architecture, ETH Zürich, Switzerland

Received 15 May 2025; received in revised form 23 July 2025; accepted 7 August 2025

KEYWORDS
Digital architecture;
Architectonics;
Artificial intelligence;
Latent space;
LLMs;
SOMs

Abstract This paper introduces CLIPedia, an entirely local, highly scalable, multimodal search engine that integrates the structured knowledge of Wikipedia into the latent embedding space of OpenCLIP. Alongside its technical development, the paper offers a theoretical framing, positing AI not merely as a tool for information access but as an architectonic instrument for invention and discovery. Built on a two-tiered Self-Organizing Map (SOM) architecture—comprising toroidal and ring-shaped layers—CLIPedia organizes over 30 million data points for fast unimodal and cross-modal retrieval. It achieves sub-second response times on standard hardware with minimal working memory footprint, delivering local performance comparable to cloud-based vector search systems and excelling on queries that return many relevant results. Beyond queries, CLIPedia enables latent journeys—termed quests—across high-dimensional embedding space. For this, the paper introduces a set of navigational metaphors and computational mechanisms—termed Orthodromes, Diadromes, Archidromes, and Thelodromes—to trace both linear and non-linear trajectories through the latent space of digitally encoded encyclopedic knowledge. The paper is accompanied by an open-source code repository.

E-mail address: anickl@ethz.ch.
Peer review under the responsibility of Southeast University.

https://doi.org/10.1016/j.foar.2025.08.003

1. Introduction

Present-day AI, particularly Large Language Models (LLMs) and services like ChatGPT, has transformed access to general knowledge. Yet while deep neural networks internalize knowledge on a massive scale, this new encyclopedia remains latent—hidden behind sleek chatbot interfaces. This calls for complementary approaches that do not compete with the scale or resource demands of foundation models, but instead focus on devising lightweight abstractions. Architecture, situated between applied and abstract knowledge, offers a valuable lens for such an approach, especially since the concrete art of applying building knowledge has long been associated with architectonics, the abstract art of knowledge building. As a new kind of encyclopedia emerges—shifting from the structured and indexical to the probabilistic and latent—architects are invited to engage with the high-dimensional spaces in which AI models operate, even when their dimensions exceed those familiar from CAD.

In response, we present CLIPedia: an entirely local, low-cost, open-source, competitively fast, and highly scalable multimodal instrument for exploring the embedding space of foundation models. In the following Theory section, its experimental design is framed by architectonic thought—positing the encyclopedia as a geodesic section within an ocean of knowledge and prompting us to imagine new instruments for its traversal.

In the Method section, we focus on the technical implementation of our thought models, positioning Self-Organizing Maps (SOMs) as meta-models to foundation model embeddings. These offer an interpretable way to chart latent space—in contrast to the black-box optimization focus of mainstream Approximate Nearest Neighbor (ANN) methods. CLIPedia consists of two layers of SOMs, trained on a large-scale corpus of Wikipedia text and images embedded by OpenCLIP vision-language transformers. This architecture enables fast unimodal and cross-modal queries, along with specialized functions that return comprehensive reference lists or image results based on textual mentions.

Still, because similarity metrics in high-dimensional models often act as a contextual funnel, narrowing scope and collapsing questions into answers, we argue that point-based queries alone remain insufficient for meaningfully exploring latent space. To address this, CLIPedia introduces quests: experiential ventures into the vast, sparse, and often counterintuitive latent spaces of AI embeddings, made navigable through conceptual and computational aids: *Orthodromes*, charting meridians between antipodes from a given starting point; *Archidromes*, which follow paths shaped by SOMs trained on external corpora acting as guides; *Diadromes*, non-linear trajectories between two embeddings interpolated by SOM topology; and *Thelodromes*, self-directed journeys guided by a SOM-based compass. These modes may be combined, enabling augmented forms of discovery.

The paper concludes by evaluating CLIPedia's performance, discussing its current limitations and potential applications—from indexing niche datasets, including those in architectural research and education, to integration with Retrieval-Augmented Generation (RAG) setups that support conversational interfaces—and reflecting on architecture as a high-dimensional art articulated in low dimensions. An open-source code repository accompanies this work, supporting reproducibility through a demo dataset and providing functions for readers to build their own CLIPedia.[2]

2. Theory

2.1. Buildings of knowledge

Architectonics is the abstraction of the art of architecture, "and holds a very similar meaning with regard to the structure of human knowledge" (Lambert, 1965, p. XXVIII). Here, the edifice is intellectual rather than material, more constellation than construction. Instead of raising a fixed structure, it becomes an "art of building active states" (Bühlmann, 2020, p. 170). Knowing that mathematics can hardly have absolute foundations, and that systems cannot be both complete and consistent (Gödel, 1931; Popper, 1998; Wittgenstein, 1976), frees architectonics from the burden of claiming ground, constructing systematizations, and raising ontologies. It is this stance that will later allow us to abstract from CLIP as a foundation model and Wikipedia as common ground—and to construct CLIPedia more as a model than a system, more as a topography than a taxonomy. By building abstract frameworks composed of multiple models and assigning each a determinate content, architectonics helps us obtain an operational organon (Serres, 2021). Computers render architectonics digital, allowing us to code, saturate, and mobilize computational models and instruments to find stability in the flood of information (Hovestadt, 2015)—entrusting, like ancient architects, that the most divergent domains of knowledge share correspondence, compatibility, and perhaps even commonality through analogy (Vitruvius, 2006). Code, always mathematical (Serres, 2022), is how we put our thought edifices to work and into play. Rather than programming in procedural, low-level, imperative languages that make things work, we code in functional, high-level, and declarative languages that allow us to think things through (Hovestadt, 2020). Mathematica invites us to do just that—offering not only computational libraries but an entire computational encyclopedia to draw from (Wolfram, 2011).

2.2. Plots of the encyclopedic circle

Any encyclopedia, a term etymologically denoting a circle of doctrine or a round of knowledge (OED, n.d.), presents a particular architectonic project in itself, famously realized through the Encyclopédie. Highlighting its cosmo-political

[2] An open-source repository is provided under https://github.com/AgostinoNickl/clipedia. It contains the complete Mathematica code for reproducing CLIPedia inference results using a sample from the pre-processed corpus and a pre-trained model. It also includes a notebook manual and specialized functions to support building a CLIPedia from Wikipedia dumps.

dimension, Simondon notes that it contains "high-level knowledge, but despite this, it is meant for all; the cost of the book alone limits the possible purchases. … For the first time, one sees a technical universe constituting itself, a cosmos wherein everything is related to everything else rather than being jealously guarded by a guild" (Simondon, 2017, p. 110). Yet, as Serres contends, the encyclopedia remains "a chaos of attempts [which] have been left at an analogous level of completion-incompletion" (Serres, 1982, p. 552). While Leibniz described the horizon of human knowledge itself as finite—because the combinations of language are finite (Forman, 2018; Leibniz, 2017)—Serres describes the non-finite relation between the encyclopedic round of knowledge and the love for wisdom in a geodesic image: "Wherever I am in the encyclopedia I find the whole of philosophy, present and profiled; wherever I am in philosophy, I find the encyclopaedic cycle in a certain latitude" (Serres, 1982, pp. 640, 641). Such a nautical image challenges the taxonomic and classificatory endeavors of organizing human knowledge, taking "the entire body of the science … as an ocean, continuous everywhere, without interruption or break" (Leibniz, 1903, p. 530; cited in Selcer, 2007).

While philosophers have long leveraged nautical and geodesic metaphors—oceans and horizons, latitudes and longitudes—to imagine encyclopedic space, these images now find renewed resonance: learning machines embed any input as normalized unit vectors based on what they have learned (Jurafsky and Martin, 2025). Borrowing CLIP's hypersphere and Wikipedia's content, CLIPedia combines two types of encoded knowledge. Separated by nearly two decades, they reflect an architectonic shift: from an encyclopedia organized around structures, classifications, taxonomies, and ontologies toward one that is non- or only latently structured—a probabilistic model that mobilizes intelligence by inductively tracing relationships across big data.

2.3. Structures of the web and wikipedia

The web emerged as a vast, open, lateral medium hosting immense volumes of multimodal content—explosive in both scale and speed of access (Virilio, 2005). Suddenly, we held the world's knowledge in our hands: Wikipedia in a click, rather than the *Encyclopædia Britannica* in a crate (Serres, 2015). Despite the protocols and infrastructures it inherited since its inception at CERN (Berners-Lee, 1989) and which enable its spread (Eco, 2009; Plant, 1997; Roman, 2021), the web refuses to be centrally controlled, mapped, or measured (de Kunder, n.d.; Plant, 1997). It is naturally abundant, scarcities artificial (Bratton, 2015), and at the disposal of anybody connected to it (Roman, 2021), even if its quantities are insurmountable without code and computation (Marinčić, 2019).

Despite occupying only a tiny fraction of the Web, Wikipedia remains by far the largest online encyclopedia (Jemielniak, 2019). Co-authored, open-source, well-referenced, and hierarchically structured, it resembles a vast network graph composed of interconnected links, formalized rules, and taxonomies—an attempt to capture "the sum of all human knowledge" (Wales, 2004). Similar to classic search engines, which match query text to hierarchically constructed or ranked indexes, Wikipedia allows exploration of its graph-based layout through hard-coded links, with text as its central modality. CLIPedia borrows the text and image content of the entire 2024 English Wikipedia. While standing in for any structured data source, Wikipedia exemplifies the encyclopedic spirit of the Web prior to the rise of LLMs. Over the past few decades, the web has evolved into a vast, multimodal repository of knowledge—not only for human learning but also for machine learning, as massive web corpora comprising up to 250 billion web pages (Common Crawl, n.d.) now form the substrate for the self-supervised training of LLMs (Brown et al., 2020; Jurafsky and Martin, 2025).

2.4. Spaces of AI and CLIP

Text transformers, such as those underpinning LLMs, are composed of multiple neural layers that learn probabilistic relationships from their training data. Operating within high-dimensional embedding spaces, they estimate likely continuations by comparing their vectorized representations to input sequences and their generated outputs (Jurafsky and Martin, 2025).

As language modeling has progressed—from statistical to neural networks (Bengio et al., 2003), and from shallow to deep architectures (Goodfellow et al., 2016)—embeddings have evolved from interpretable, static term-frequency vectors to contextually fluid, transformer-based representations (Jurafsky and Martin, 2025). Encoded as distributions of linguistic probability and harvested at a scale far beyond what structured repositories like Wikipedia could offer, this knowledge remains largely inaccessible, implicitly hidden within the weights and biases of massive models. While well-crafted prompts may bring fragments of it to light (Nickl and Bokhari, 2023; Bokhari and Nickl, 2024), the ties between their responses and training references are currently cut for good (Akyürek et al., 2022). Whether probabilistically inferred or plausibly hallucinated (Bubeck et al., 2023; Maynez et al., 2020), this artificially intelligent encyclopedia presents an answer to every question, without disclosing the knowledge it draws from.

Multimodal models allow for queries posed in either images or text to be answered with either text or image — through closest-match retrieval, as in CLIP, or outright generation, as in BLIP (BLIP-2; Li et al., 2023). Using a vision and a text transformer, CLIP does not generate text or images, but positions them precisely within its learned, shared embedding space (CLIP Multi-domain Feature Extractor; Radford et al., 2021). CLIPedia builds on OpenCLIP Laion 2B, an AI model pre-trained on two billion image-caption pairs from the web (Cherti et al., 2023; OpenCLIP Multi-domain Feature Extractor, 2023), likely including some Wikipedia content, given that the latter can only account for a small fraction of the training corpus.[3]

[3] Wikimedia Commons hosts approximately 100 million media files, only a portion of which are captioned images—amounting to less than 5 % of the 2 billion image-caption pairs in LAION-2B ("Wikimedia Commons," n.d.; LAION, n.d.).

Before setting out to explore, as Serres puts it, "a few cantons in Encyclopedia country" (Serres, 1982, p. 536), as laid out by AI, we need charts and navigational instruments.

2.5. Mechanics for models and maps

Through encoding, we can map all of Wikipedia's structured content, image or text, into the un- or only latently structured embedding space of OpenCLIP—each embedded element becoming a concrete landmark in it. Metrics now allow architects to play mathematical games—not only by measuring distances between towers and castles (Alberti, 2010), but between encoded topics and concepts. In contrast to the matrices that derive and the networks that generate them, this space is not a graph nor a system; there are no nodes, no edges, and no fixed rules; Relations are not formalized as hyperlinks—here, everything is related to everything to differing degrees. If we were to pose a question and project it into this space, we could find a response by seeking its nearest neighbor. To do so efficiently, the shards of once-structured encyclopedic knowledge need to be reorganized—but instead of reorganizing them through systematization, we can let their embeddings organize themselves, taking SOMs as models (Kohonen, 1982; Hovestadt, 2015).

As an algorithm, its adaptability has already given rise to a myriad of architectonic instruments and architectural tools (Alvarez Marin, 2020; Cai et al., 2024; Marinčić, 2019; Orozco, 2017; Roman, 2021; Saldaña Ochoa, 2021; Zaghloul, 2017; Zifeng, 2021). As a map, it allows us to chart and navigate a world along the latent lines tied by AI. As a model, it enables us to effectively measure distances to encoded landmarks by providing synthetic markers. As a meta-model to any embedding model, SOMs offer a navigable topology of how data is embedded; like a spectrum revealing the hidden composition of light, they highlight the latent distribution of data at any chosen resolution.

By mapping encoded elements onto their trained spectrum, SOMs not only retain the imprint of their training in abstract weights but can also incorporate concrete data points—becoming quasi-maps for the quasi-territories of a "digital continent" (Matthias Boeckl, cited in: Bühlmann, 2024, p. 42). As latent affinities settle into topological relations, qualities submerged in quantity come to the surface, and formerly dispersed clusters turn to connected landscapes—neighborhoods of meaning that architects can learn to explore: not by walking, as in Boston (Lynch, 1960), nor drifting, as through Paris (Debord, 2007), nor driving, as in Las Vegas (Venturi et al., 1988), but by inventing new modes of querying and questing.

2.6. Instruments for queries and quests

We can query CLIPedia through any text or image — its embedding casts us into the multicursal maze of our map (cf. Aarseth, 2012). We may gain some orientation by looking at the text and images closest to us. But while an intelligent search engine has its uses, our queries limit us to places we already know how to address. Quests, however, free us from retracing what we already know and invite us architects onto computational grand tours. Inspired by old navigational metaphors of encyclopedic space, we propose four instruments that chart distinct paths through latent space.

As we might intuit that the embedding 511-sphere of OpenCLIP is best explored along its seams—much like the 2-sphere we call home—we can think of *Orthodromes* as journeys along geodesic great-circle segments between two points. But just as flying over continents teaches us little about the land below (Serres, 2018), we learn next to nothing by traveling along the shortest path over a digital continent (see 3.4.1). Our encoded landmarks reveal a landscape that is non-uniform, to be thought of less as a landmass than as a *vectorial archipelago*—a space akin to densely packed forested islands amid endless stretches of nothingness, like a dead calm sea. To make our journeys worthwhile, we must move non-linearly through it.

Archidromes are guided journeys through CLIPedia, led by those who have gone first. Just as woodcutters cultivating the woods know best how to navigate them (Heidegger, 2002), so might authors cultivating the word know best how to guide us through a latent space co-constituted by text—even though selecting those as guides who witnessed the emergence of quantum mechanics and computation requires special scholarly care.[4] Where specific and general bodies of knowledge intersect, we establish points of mutual contact, allowing for anachronistic communication between past and present, and analogical conversations across regions of the encyclopedia. *Diadromes* are interpolated journeys between two locations when there are no guides to depend on—probing, sampling, and scouting a way through the high-dimensional CLIP space rather than imposing structure upon it. *Thelodromes* are self-guided journeys, informed by a spectral compass. Since from any given standpoint we are likely not to see the forest for the trees, it provides us with the principal directions along which we can choose to move. All these instruments may be combined: an Archidrome might be interpolated by Diadromes, a Thelodrome weighed by Archidromes.

Whichever instrument we choose, our interfaces as traveling architects are not pages in a paper sketchbook, but cells in a computational notebook — our queries and quests unfold in *QueryBooks* and *QuestBooks*, hypertextual readers and visual essays on any imaginable topic. In the following section, we will discuss how to build CLIPedia and mobilize it with a set of instruments—not only as thought models, but as computational models.

[4] Engaging with the work of authors who died less than 75 years ago often involves working with copyrighted material, even if on a small scale—such as excerpts from individual books. When conducted for academic research and aimed at revisiting original statements in contemporary contexts without reproducing substantial portions of the source material, such use is consistent with fair use principles. In our setup, the SOM can offer internal indices and relate them to a generic, open-source corpus in a way that supports transformative analysis, unlike generative language models that generate new or synthetic content. If correct metadata is provided, the system automatically compiles citations.

3. Method

Having situated CLIPedia within a theoretical frame, we now turn to the technical implementation of a highly generalizable "informational instrument" (Roman, 2021, pp. 290, 291)—one that offers a scalable architecture for fast, vector-based search of embedded knowledge. We collected data between May and August 2024 on a single Windows 11 workstation (Intel Core i9-12900KF, 64 GB RAM, NVIDIA RTX 3080 Ti). Processing included parsing the April 2024 English Wikipedia dump into XML over ~4 days, converting Wikitext to cleaned text over ~7 days, extracting and encoding ~30 million paragraphs over ~9 days, and collecting and encoding ~2.2 million images over ~57 days. In total, our workflow (Fig. 1) ran continuously for 11 weeks, and we monitored it remotely to ensure stability and resolve occasional interruptions, facilitated by robust routines and logs. We performed all steps using Wolfram Mathematica scripts, which we have openly provided in a dedicated public repository on Github (Nickl, 2025).

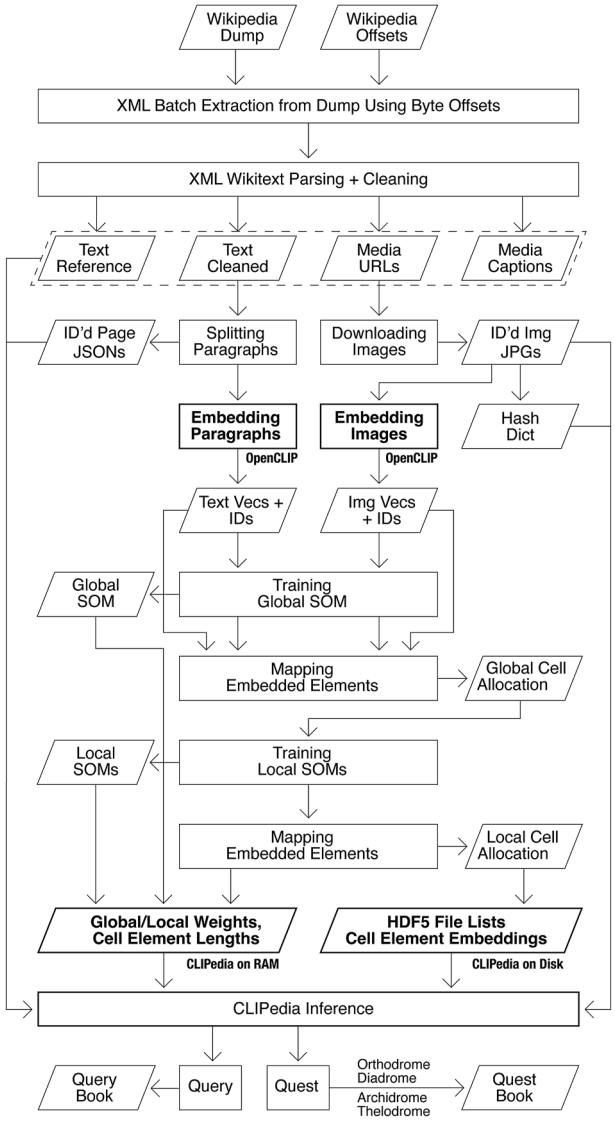

Fig. 1 The data processing steps of CLIPedia, from Wikipedia Dumps to QueryBooks and QuestBooks.

3.1. Sourcing CLIPedia

3.1.1. Parsing data

Abundant and permissively licensed, Wikipedia offers us an ideal source for image and text data. As parsing from a local file is generally faster than crawling from the web, periodic dumps are our starting point—each a compressed, dated archive of all English Wikipedia pages, consisting of a 25 GB compressed XML file and a small index file.[5] The latter allows for byte ranges corresponding to batches of ~100 pages in Wikitext format to be loaded incrementally into memory, decompressed, and converted into 236,481 individual files. For the articles contained in each, we look for markers indicating pages that disambiguate, redirect, categorize, or list. These are discarded along with internal documentation, incomplete drafts, and stubs. For all others, we split the main text from the sources section, clean the Wikitext using a custom function containing hundreds of regular expressions and string patterns, designed to robustly remove classes, templates, and formatting tags. In-text references and image links with their captions are extracted and replaced with numbered placeholders. Each of the 6,470,372 processed pages is then saved as key-value pairs, containing text, references, sources, image links, and captions. Only now do we begin to access all image links and attempt downloading associated files in JPG format. To avoid duplicates of figures shared across articles, images are saved under a truncated, MD5-hashed link.

3.1.2. Identifying data

For further processing, pages are discretised into paragraphs, commensurate in length with the text sequences that AI models are trained on, which range from single sentences—like captions in CLIP,[6] to several paragraphs in current LLMs.[7] Paragraphs are extracted between natural breakpoints such as double returns and further divided if they fall below a threshold; a custom function then merges or splits sections into sub-paragraphs of ~200–500 characters each, approaching the upper bound of CLIP's token input limit. Since the Mathematica implementation of OpenCLIP handles longer inputs gracefully, we accept this compromise to preserve as many natural paragraphs as possible as self-contained units of encyclopedic knowledge. With all texts split and all images downloaded, we have compiled our Wikipedic Corpus: 30′330′012 texts and 2′239′112 images. Each of them is given a unique identifier; images retain their hashed URLs, texts are identified by page, paragraph number, and their position within if split.

[5] CLIPedia works with the data from the 20. April 2024 dump, downloaded from the official dump repository (Wikimedia Foundation, n.d.).

[6] Max. sequence length of CLIP is stated as 76 tokens (Radford et al., 2021, p. 5), 1 token ≈0.75 words in English, ca. 57 words.

[7] Max. context windows of Llama: 2048, Llama 2: 4096 tokens, 1 token ≈0.75 words in English, ca. 1500/3000 words (Touvron et al., 2023, p. 47).

These identifiers do more than facilitate retrieval by serving as filenames—they also encode additional information: image identifiers point to their web sources, and page handles begin with their Wikipedia Page ID, enabling straightforward URL reconstruction.

3.1.3. Encoding data

OpenCLIP has been pre-trained on 2 billion captioned images (Cherti et al., 2023; "OpenCLIP Multi-domain Feature Extractor," 2023), making it a robust choice for encoding our Wikipedic Corpus. While more advanced models, such as JINA-CLIP, have become available since (Jina AI, 2024; Koukounas et al., 2024), OpenCLIP remains a widely distributed model, with CLIP variants still occupying the top of the list of downloads of model platforms like Huggingface.[8] We let its two complementary models (both using the ViT-B/32 specification) generate embeddings for all text and images contained in our corpus, leveraging the processing power of our GPU. To optimize storage, memory, and computation time when dealing with tens of millions of encoded elements, we store each as a simple list, consisting of a string as its persistent identifier and a list of floats for its scalars. By computing cosine distances between these embeddings, we can now sort them based on similarity for any query we embed.

3.2. Building CLIPedia

3.2.1. SOMs as meta-models

Conventional systems typically use ANN algorithms to find the closest corresponding element within an embedded corpus. These are often implemented via Hierarchical Navigable Small World (HNSW) graphs or Inverted File Indexes (IVF), as employed by Facebook AI Similarity Search (FAISS), and are widely deployed in cloud-based vector databases, where their high memory demands are less of a constraint than on standard consumer hardware (Mazanec and Hamzaoui, 2022). The Vamana algorithm, used in DiskANN, delivers strong performance using SSDs and modest RAM, although it requires low-level tuning of soft- and hardware, and is unsupported on macOS (Subramanya et al., 2019; Microsoft, 2025). While these algorithms are engineered for raw efficiency, they produce topologically disordered maps that offer little in terms of interpretability, even when applied to two-dimensional data (cf. Subramanya et al., 2019).

In contrast, SOMs draw charts that aim to preserve topographic relationships in high-dimensional spaces (Kohonen, 2001a). They reduce dimensionality non-linearly while maintaining topological continuity across cells (Sammut and Webb, 2017), and approximate the distribution of the input space (Kohonen, 2001a). Graphs of connected nodes—vectors matching the dimensionality of the training data—are initialized with randomly assigned weights (Kohonen, 2001b). Over a series of iterations, these so-called neurons compete to become Best Matching Units (BMUs) with the least Euclidean distance for randomly sampled input vectors (Kohonen, 2001b).

Together with their neighboring neurons, they adjust their weights based on a Gaussian neighborhood function (Kohonen, 2001b), which shrinks its radius from global to local for each iteration, modulating the plasticity of the neural net (Kohonen, 2001b). SOMs can be used as pre-trained models, as is the case with CLIPedia, or trained on the fly, serving as probabilistic instruments for their own traversal, as shown in 3.4.

3.2.2. Model architecture

The size and topology of SOMs are set before training. A large grid disperses data into smooth distributions, while a small grid concentrates data into sharp clusters. Seeking high-resolution maps, we opt for larger grids; CLIPedia measures 32×32 cells at its global level (Figs. 2 and 3). It is initialized as a two-dimensional grid without boundaries, allowing for periodicity in both dimensions—cells located at one edge of the map connect seamlessly to their opposite side (Kohonen, 2001b). This mitigates edge effects common in SOMs, which distort distributions along the boundaries due to their cells having fewer neighbors (Mount and Weaver, 2011).

During training, our batch algorithm (ETHZ CAAD et al., 2020; Marincic, 2021) updates all neurons simultaneously rather than sequentially (Kohonen, 2001b), which allows us to precompute packages of training samples containing the same number of encoded image and text vectors for each iteration, without needing to keep all of them in memory while training.[9] Rather than decaying linearly, our neighborhood function follows an exponential decay, promoting faster convergence at the risk of overfitting. To assess quality, we measure Quantization Error (QE)—the average distance between input vectors and their closest BMUs—and Topographic Error (TE)—the proportion of input vectors for which the first and second BMUs are not neighbors (Kohonen, 2001b). After 32 training iterations, we find the optimum at iteration 17, where QE has already converged and TE remains low, as indicated by error curves and two-dimensional renderings, which reduce weights to RGB color space using PCA or t-SNE (Fig. 2).

3.2.3. Distributed storage

While training on the Wikipedic Corpus yields a lightweight map, applying its full content afterward would place too heavy a load on working memory. To manage memory efficiently, the embeddings associated with each SOM cell are stored locally in bundles, as their self-similarity makes them the most relevant candidates for local comparison during downstream use. Because the tens of thousands of elements in each cell of our SOM remain too varied and voluminous to be handled in batches, we must introduce a second level of granularity.

For each of the 1024 cells in the global, two-dimensional toroidal SOM, we instantiate a layer of local, one-dimensional circle SOMs—mapping not a region but a sequence (Kohonen, 2001b). The number of elements contained in each cell automatically determines the initial size

[8] openai/clip-vit-base-patch32 still occupies the 12th rank on Huggingface, as of 18th July 2025 ("Models - Hugging Face," 2025).

[9] In CLIPedia's case, each training batch was equally split in image and text encodings, roughly containing up to 3 % of the whole dataset for each iteration.

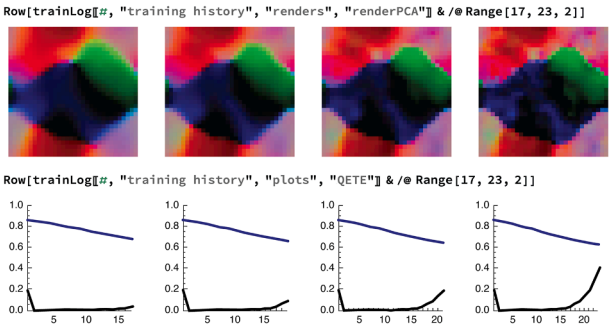

Fig. 2 A range of SOM training iterations rendered in PCA, below a combined graph showing QE and TE.

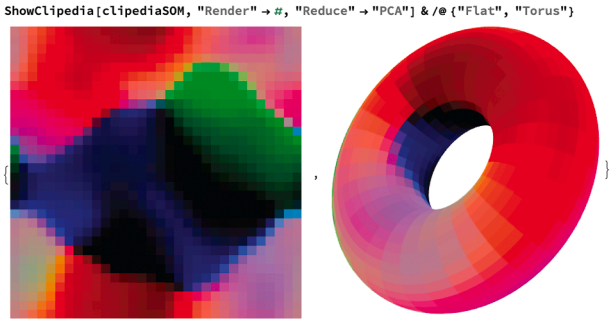

Fig. 3 Left, CLIPedia's unrolled first layer, showing the "mouth" containing texts. Right, CLIPedia as a torus.

of its local SOM, which they are then mapped onto.[10] After training for 15 iterations, the best model is automatically selected based on QE and TE. The result is 32,324 sorted batches of highly similar elements; the SOMs thus serve as an index for intelligently distributed storage, with all vectorized elements locally accessible as HDF5 files. Their original counterparts reside in hashed folder structures: a text folder containing page-level JSON dictionaries, each keyed by paragraph ID, storing all individual paragraphs; a page folder with dictionaries recording each page's title, references, and image mentions; an image metadata folder listing each image's Page IDs and associated captions; and finally, the images themselves (Fig. 1).

Taken together, this architecture renders CLIPedia exceptionally lightweight: its limits are determined not by working memory but solely by local storage. All models, data, and functions occupy ∼430 GB of disk space, while the memory footprint remains minimal—consisting mainly of a 218 MB data structure containing the dimensions, weights, and counts of image and text encodings for each locally stored SOM cell.

3.2.4. Multi-modal gap

Unfolding the global, toroidal layer of CLIPedia and rendering it in PCA reveals two distinct regions: one dark and mouth-shaped, containing almost exclusively text encodings, and another surrounding it, inhabited almost exclusively by image encodings (Fig. 3). While the oral shape is coincidental, the rift is not: it is the expression of CLIP's "modality gap" (Liang et al., 2022, p. 1).

Despite being embedded into a learned multi-modal space, encoded texts and images remain distinct, their original domains latently present in their distribution. As "the only interaction in a CLIP model between the image and text domain is a single dot product in a learned joint embedding space" (Radford et al., 2021, p. 27), this is hardly surprising. While CLIP's two constituent transformers learn to share the same field and train towards the same goal, they do not play the same game: depending on their input domains, their embeddings inhabit entirely different regions of the shared latent space, residing on cone-shaped, lower-dimensional manifolds within it (Liang et al., 2022).

Whether this separation results from CLIP using the dot product as its main learning objective and attending solely to relations within and not across each domain (Radford et al., 2021) or whether the initialisation method, the contrastive principle, or deep neural nets themselves are responsible (Liang et al., 2022): as long as the models perform well despite this, researchers see no reason to bridge this gap (Liang et al., 2022). To us, it serves as a reminder that any embedding is contingent on the model performing it, and that CLIPedia's embedding space is necessarily wonky, biased, nonlinear, and non-uniform.

3.3. Querying CLIPedia

3.3.1. Retrieval mechanism

With CLIPedia's global and local layers in place, we can now retrieve elements—the lower the distance, the higher the similarity (Jurafsky and Martin, 2025), and the bigger the corpus size, the closer a corpus may come to offering an approximate decoding for any encoded input. Euclidean and cosine distances—standard metrics for vector comparison (Qian et al., 2004)—respectively capture positional and angular differences (France et al., 2012; Jurafsky and Martin, 2025). Already an integral part of CLIP's training (Radford et al., 2021), cosine distance is used to measure similarity within CLIPedia.

Given any CLIP-embedded text or image, we compute the similarities to the 32,324 cells in CLIPedia's second layer and sort them into a list of BMUs. Since distances reduce complex, high-dimensional relationships to a single value, multiple cells may contain relevant matches. From the top-ranked BMUs, we identify how many local batches of encoded elements must be loaded into memory to reach our Ranked Corpus Sampling (RCS) parameter, the fraction of the image or text corpus considered in response to a query—this parameter allows for tuning the trade-off between speed and accuracy. Once selected, the relevant elements are loaded into memory and locally sorted by similarity across CPU cores. The results are then aggregated and globally ranked into a unified list of Best Matching Elements (BMEs).

[10] As the size of the SOM, we take the maximum between 2 and the rounded number of elements in the cell (pertaining to images or texts) divided by 1000—if elements are too little, we use a preset data structure of a SOM with one cell only for consistency.

3.3.2. Query functions

AskClipedia (Fig. 4) is the key function for queries: it accepts text or image input and allows us to specify our preferred modality for outputs and their number, providing answers in well under a second (see Section 4.1). We can search for text by text, image by image, from image to text, or from text to image, asking for 10, 100, or 1000 results (Figs. 5 and 6). The natural language capabilities of OpenCLIP allow us to query with the nuance known from LLMs (Fig. 1), rather than relying on structured query languages or term-based search engines. Image queries, similarly, go beyond mere visual similarity; they carry the rich contextual associations acquired through contrastive learning.

Other bespoke query functions enable advanced searches—such as tuning the RCS parameter to screen a higher percentage of the corpus, compiling comprehensive bibliographies by extracting in-text references from results for a specific query (Fig. 7), or retrieving images not by proximity but through textual embeddings—specifically, their mentions in paragraphs (Fig. 8). Whether image or text, results are laid out in output cells that, thanks to their identifiers, are hyperlinked to their respective articles online, and that disclose references or captions as tooltips.

Already stored in our notebooks, they are ready for further computation or export as *QueryBooks*. Building on these retrieval mechanisms, we now turn to the exploratory methods that let us navigate CLIPedia's latent space.

3.4. Questing CLIPedia

3.3.1. Orthodromes

Orthodromes move along geodesic circle segments connecting two points. OpenCLIP embeddings, having unit length 1, lie on the surface of a unit hypersphere. Taking any vector as a starting point (e.g., the embedded string "Column Capital"), we obtain its antipode by multiplying it by −1. Interpolating between these poles yields intermediate points. While spherical linear interpolation (SLERP), commonly used in CLIP spaces (Ramesh et al., 2022), fails on antipodes due to the infinite number of possible paths, linear interpolation (LERP) provides a set of intermediate vectors (e.g., 16 steps), though these do not strictly adhere to the hypersphere. Each intermediate point is approximately decoded by retrieving its nearest neighbor.

Notably, linear interpolation leads to polarization: cosine distances shift from ∼0 near the start to ∼2 near the antipode, likely retrieving only the two antipodal

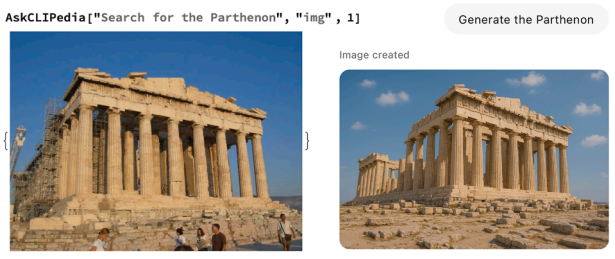

Fig. 4 Left, the top-ranked result in CLIPedia; Right, a generated response by DALL·E−3 via ChatGPT in 2025.

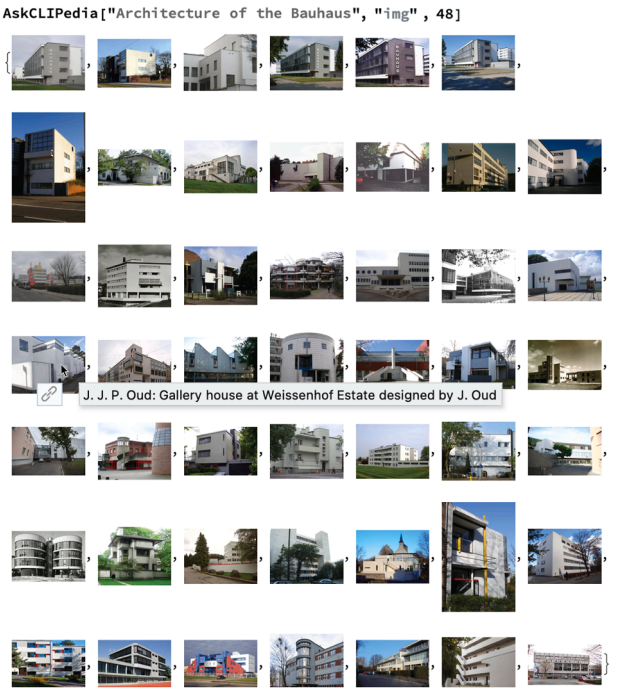

Fig. 5 A text query for the 48 closest images, all hyperlinked to their source and augmented with captions.

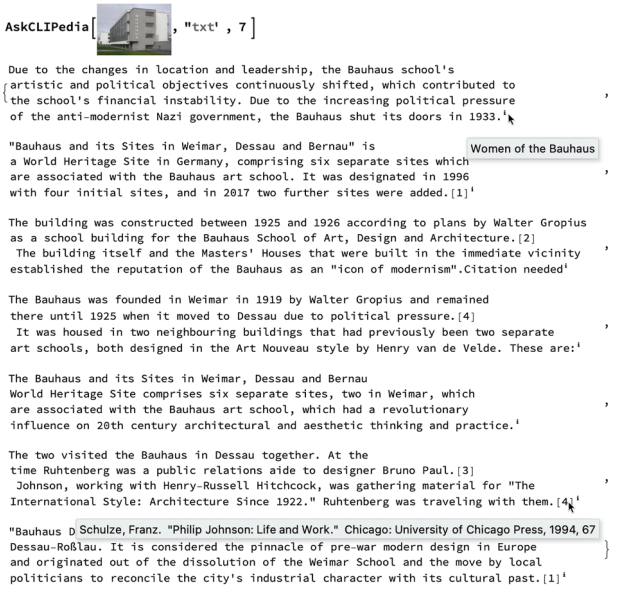

Fig. 6 An image query to the seven closest paragraphs, all results being hyperlinked and referenced.

results (Fig. 9). To improve meaningful traversal, we implement an Orthodrome function that uses a third, disambiguating vector (e.g., the embedded string "Frieze") to plot a great-circle segment between antipodes. Along this path, distances increase smoothly along a sigmoid curve, and intermediate matches emerge more consistently. Orthodromes, as geodesics on the hypersphere, nevertheless cover only limited space (see Section 2.6); even with only 8 interpolation points, and drawing from a

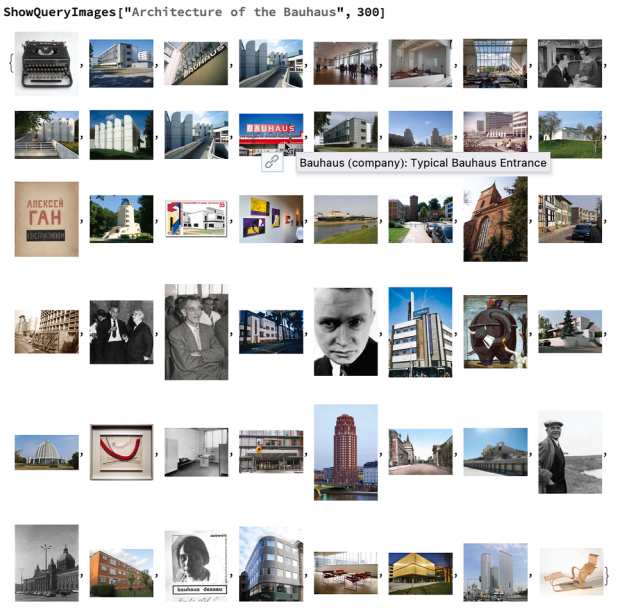

Fig. 7 A list of image references for a query, compiled from image links embedded in the 300 closest results.

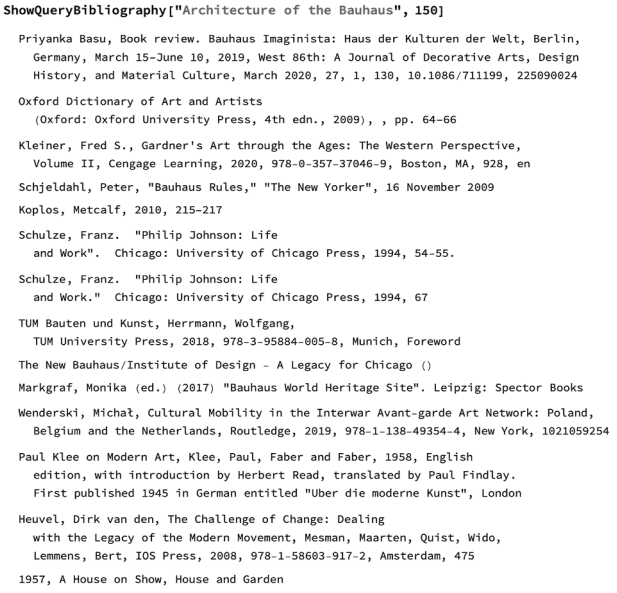

Fig. 8 A list of bibliographic references for a query, compiled from references embedded in 150 results.

corpus of several million elements, the number of distinct matches remains low (Fig. 10).

3.4.2. Archidromes

Archidromes are guided movements through CLIPedia's latent space. To take selected authors as our guides, we first process their texts. As these are often books, a bespoke function records metadata for each input PDF file for future reference and attribution, while another function cleans page content by extracting repeating substrings at the top and bottom of each page, which are likely book or chapter titles. Once split into paragraphs and encoded

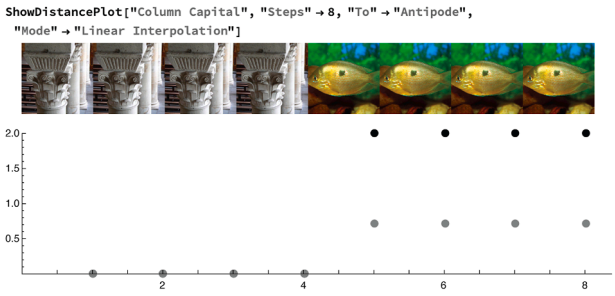

Fig. 9 Linear interpolation in 8 steps; results and distances (in grey: closest match) show stark polarization.

into OpenCLIP, we train guide models as periodic, one-dimensional SOMs with a default of 16 cells (Fig. 11). Instead of generating a single query point, the guide SOM produces multiple reference points covering as much of OpenCLIP's 512-dimensional space as possible, together offering a characteristic perspective on CLIPedia.

After specifying whether to retrieve images, texts, or both, we compute BMEs in the encoded guide corpus for each guide SOM cell and identify the corresponding BMEs in CLIPedia. If several guide cells align with a single CLIPedia cell, the matches are aggregated. As we move along the guide spectrum through different regions of CLIPedia, query results are compiled into notebook cells under each guide BME, displayed in bold with automatic page attributions. Since each cell may contain multiple nearby matches, navigation controls allow for interactive exploration of neighboring results (Fig. 13, left).

3.4.3. Diadromes

Diadromes move along non-linear interpolations between two embeddings. Taking both as query vectors, we first retrieve a sufficiently large set of BMEs (typically 100 or 1000), assuming this subcorpus contains enough results to pave a probabilistic path between departure and destination.

We train a one-dimensional, circular SOM on this subset, and map the closest encoded elements to its cells (Fig. 12). Unlike an Orthodrome, which traces a purely geometric circle on a sphere, the Diadrome produces a topological circle that adapts during training to the concrete content distribution in the embedding space. The circular SOM yields two alternative paths, separated by the cells nearest to the departure and destination points. This approach can

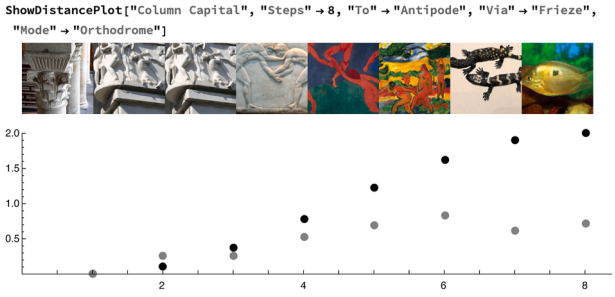

Fig. 10 Orthodrome in 8 steps, showing varied results and smooth interpolation of distance (in grey: closest match).

Fig. 11 Archidrome of 16 steps, moving along the SOM spectrum trained on the text of an author, colored by PCA.

Fig. 12 Diadrome of 18 steps, from a SOM interpolating between points by organizing results obtained from both ends.

also be used to compute non-linear connections between the spectral positions of an Archidrome, augmenting it with contextual content and smoothing out abrupt jumps (Fig. 13, left).

3.4.4. Thelodromes

Thelodromes are self-chosen movements in latent space, assisted by a probabilistic compass. Any string or image can serve as a starting point, from which we query the surrounding region. With a query scope of 100 or 1000 results, we dynamically train a circular, one-dimensional SOM to compute principal directions. While we can restrict these to four, we can set as many as needed. After less than a second's worth of computation, we obtain characteristic texts or images pointing towards each direction, which we can inspect interactively in a QuestBook.

When selecting a direction using a *NextStep* function, the current position is updated and stored as a global variable. This becomes the next query for the compass. At any time, we can limit the search to text, images, or both using toggles in a pop-up window (Fig. 13, bottom left). Optionally, the compass can be loaded with a guide corpus, and its suggestions weighted against the median of that corpus's encodings to bias directional suggestions (Fig. 13, right).

When a Thelodrome is thus combined with an Archidrome, directions are still computed agnostically, but one is marked as the guide's most probable preference. If a similarity threshold between any guide encodings and the user's selections is exceeded, a button appears; when clicked, matching quotes from the guide corpus are displayed in a pop-up window (Fig. 13).

4. Discussion

We have introduced a method for building a dependable, lightweight, and highly scalable architecture for organizing, storing, and exploring tens of millions of high-dimensional embeddings and their multimodal counterparts. Our two-tiered architecture—comprising a toroidal global SOM and nested ring-shaped local maps—enables efficient and adaptive navigation of latent space without the opacity and memory overhead of conventional ANN methods. CLIPedia includes a set of bespoke query functions, instrumental for querying an encyclopedia with the semantic sensitivity known from LLMs, and quest functions—Orthodromes, Archidromes, Diadromes, and Thelodromes—which serve as encyclopedic instruments for navigating sparse, nonlinear latent spaces. While the architecture of CLIPedia is not itself a foundation model, but rather an abstraction of one, it invites exploration of spaces that would otherwise remain latent; its functions enable experiences distinct from those offered by commercial chatbot interfaces. Constituted by CLIP and Wikipedia, both components can be seen as placeholders for any embedding model and multimodal corpus, making our proposed architecture a lightweight complement to the rapidly shifting landscape of AI. In the next section, we evaluate CLIPedia's performance, examine its applicability, and reflect on the role of the architect in a world newly indexed by AI.

4.1. Evaluations

To evaluate the performance of our SOM-based vector retrieval system, we compare it against contemporary ANN solutions. While differing in architecture and deployment, Microsoft's DiskANN benchmark for Azure Cosmos DB provides a useful reference point, indexing a similarly scaled corpus of 35 million vectors with 768 dimensions, also sourced from Wikipedia (Upreti et al., 2024). Our CLIPedia index occupies just 218 MB of RAM, runs entirely locally wherever Wolfram Mathematica is supported, and was tested on a MacBook Pro (M1, 64 GB RAM, 2 TB SSD), with no low-level tuning of software or hardware. Despite this off-the-shelf setup, CLIPedia's parallelized retrieval function achieves recall@10 and recall@50 rates of up to 94% and 93% (Table 2), respectively—comparable to DiskANN's reported performance (Table 1) of 90.73% and 95.67% recall.[11]

While somewhat slower in low-k retrieval of text, likely because the SOM architecture necessarily loads and scans batches rather than isolated elements, CLIPedia delivers similar latency in high-k retrieval scenarios, with as little as

[11] Recall@k measures the proportion of the true nearest neighbors found within the top-k results returned for a query vector in an approximate search setup (Upreti et al., 2024). We sampled 100 pseudorandom vectors for images and text from the indexed corpus to serve as query vectors. During evaluation of ground truth, exact and approximate (RCS = 1/0.2), with (RCS = 1), the query vector was excluded from the result set. Recall@k was computed based on the remaining top-k neighbors. Average times were evaluated with the function *AbsoluteTiming*, recording wall-clock time.

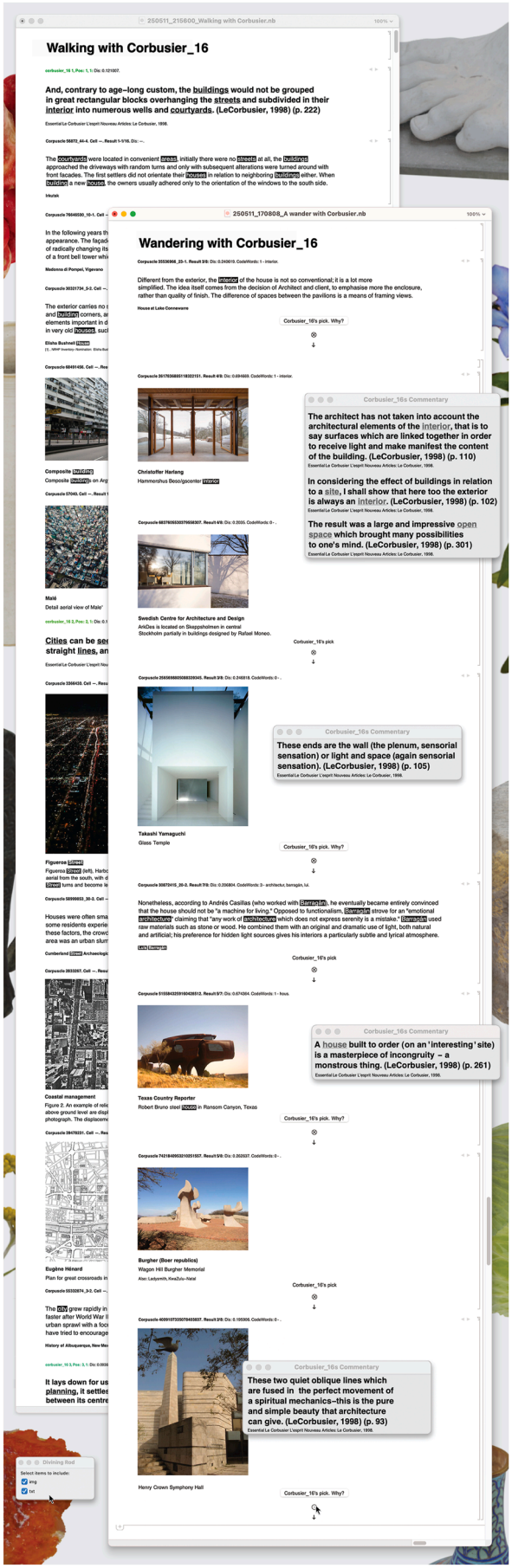

Fig. 13 Screenshots of two QuestBooks, containing an Archidrome interpolated by Diadromes (left, back), and a Thelodrome augmented by an Archidrome (right, front).

102 ms for image queries and 581 ms for text at recall@50, compared to Microsoft's reported 569 ms.[12] It further demonstrates strong recall@1, reaching 90% accuracy at just 60 ms for image queries and 340 ms for text—a metric not included in Azure's benchmarks but essential for closest-match tasks. These performance levels are tunable via the RCS parameter, which defines the fraction of the corpus subjected to close screening. While DiskANN benefits from cloud-scale infrastructure, specialized hardware, and low-level optimizations, it incurs notable latency increases as k grows. In contrast, CLIPedia maintains stable response times from $k = 10$ to $k = 50$, thanks to its pre-ranked BMU clustering. This makes it suitable for RAG and other tasks requiring multi-result retrieval.

4.2. Limitations

Since optimization lay beyond the scope of this architectonic experiment, the current CLIPedia architecture offers several opportunities for improvement. Tuning the proportion between the global and local SOM layers, as well as adjusting the dynamic size of local SOMs relative to the number of data points projected from each global cell, may yield configurations that reduce latency and improve recall. While a multimodal SOM has conceptual advantages, CLIPedia's mode-agnostic training may compromise retrieval quality, as the multimodal gap can skew data distribution across the shared SOM. Rather than combining embeddings from separate models, adopting more advanced encoding models that generate universal multimodal embeddings—such as those from the GME or MM-Embed model families (Zhang et al., 2025; Lin et al., 2025)—may help close this gap and improve overall performance, though such models are themselves still evolving. Additionally, CLIPedia currently operates near the upper token limit of OpenCLIP, which may reduce performance for long paragraphs. More recent CLIP-derived models, such as JINA-CLIP, can help address this limitation by embedding much longer text sequences and supporting more nuanced text retrieval (Jina AI, 2024; Koukounas et al., 2024).

Finally, CLIPedia is built entirely on Wikipedia data and OpenCLIP embeddings. While this ensures consistency and reproducibility, it narrows the range of possible perspectives. Expanding CLIPedia to include additional embedding models and more varied corpora would support the development of a more pluralistic encyclopedia. Since the SOM-based architecture is fundamentally content- and model-agnostic, switching to alternative or fine-tuned embedding models—effectively transforming CLIPedia into, for example, JINA-CLIPedia, BLIPedia, or GMEpedia—requires only a single change in the processing pipeline (Fig. 1). This

[12] We ran evaluations separately for text and image queries, calculating recall at $k = 1$, 10, and 50 for each RCS value. Each evaluation block was run from a cold start with a freshly initialized Wolfram kernel.

Table 1 Azure Cosmos DB Vector Search with DiskANN: 35M Vectors (Cohere Wiki-Embeddings), 768 dim. (Upreti et al., 2024).

Scenario	Avg Latency (ms) $k = 1/10/50$	Recall@k $k = 1/10/50$
35M	X/112/569	X/90.73/95.67

flexibility is a key strength of our approach, offering promising directions for future adaptation and development.

4.3. Applications

CLIPedia complements online access to knowledge by providing offline and multimodal search capabilities that extend beyond keyword-based retrieval. Unlike Google, shaped by ads, click metrics, and SEO; Wikipedia search, confined to predefined page structures; or LLM services, prone to hallucinatory responses and with latent spaces kept off-limits, CLIPedia enables cross-modal exploration through a continuous embedding space. It supports lateral discovery and serendipitous encounters, with customizable parameters and an interpretable architecture—all within a coding environment where outputs can be repurposed for diverse downstream tasks. For example, a single image of the Bauhaus building can yield paragraphs discussing the movement, relevant reference lists, visually similar images, and textually related figures, and a starting point for an open-ended exploration of the surrounding latent space.

In educational settings, CLIPedia supports studios and seminars by enabling the rapid assembly of hyperlinked, image-rich visual essays and thematic bibliographies in response to multimodal queries—offering immediate entry points into diverse fields and supporting both guided and open-ended quests.

In practice, its architecture is adaptable to domains far beyond encyclopedic content, making specialized knowledge locally searchable. Scaling with disk space rather than RAM, CLIPedia provides a blueprint for creating locally searchable, multimodal models tailored to specific needs. As interest in AI-augmented access to digital collections grows alongside concerns over the hallucinatory tendencies of LLMs (Hodel, 2024)—this architecture is well-suited for deployment in public and private archives, corporate repositories, or classified environments containing sensitive data. Running entirely locally and built on open-source models like OpenCLIP, it supports secure, vector-based retrieval without the costs or risks associated with cloud platforms, offering a viable backend for localized RAG workflows. Here, CLIPedia's topology-preserving organization facilitates fast access to thematically coherent clusters, serving as high-quality inference contexts for LLMs and supporting referenceable responses that extend the knowledge of any single foundation model.

4.4. Reflections

Apart from its applications, CLIPedia's computational implementation also invites reflection on its architectonic framing. Where encyclopedias once sought to stabilize knowledge through structure — taxonomies, classifications, hierarchies—we now mobilize it through code within latent space. CLIPedia criss-crosses all pages, topics, and hierarchies of Wikipedia; the same finite content, freed from its structure and embedded into many dimensions, becomes a liquid resource playing out in another, foreign space. Where CLIPedia retrieves the nearest index, generative AI synthesizes a new instance for each prompt. The Parthenon in CLIPedia and a probable Parthenon generated by ChatGPT illustrate how—at least along the commonplaces of the digital continent—the indexed and the inferred may converge (Fig. 4). But by carving a path between the broad footsteps of our guides, or choosing our own route with the help of a spectral compass, CLIPedia also allows us to move beyond those commonplaces and the well-trodden paths connecting them. Archidromes, Diadromes, and Thelodromes prove to be not only conceptual metaphors—inspired by Serres's encyclopedic geodesics and Leibniz's ocean of knowledge—but also computational mechanisms for exploring a vectorial archipelago.

While this may qualify CLIPedia as an architectonic instrument, it also carries architectural implications—rekindling old notions of architecture as inherently high-dimensional. Today, this dimensionality extends beyond the parameter space of cost, carbon, or other metrics (Witt, 2022), encompassing the numerous latent spaces of

Table 2 Clipedia with Double-Tier SOM: 33M Vectors (CLIP Wiki-Embeddings), 512 dim. Recall values rounded to nearest integer, standard settings in bold.

RCS: Fraction of respective corpus	Image vectors		Text vectors	
	Avg latency (ms) $k = 1/10/50$	Recall@k $k = 1/10/50$ against ground truth	Avg latency (ms) $k = 1/10/50$	Recall@k $k = 1/10/50$ against ground truth
0.0025	43/43/51	75/71/68	226/231/241	85/80/75
0.005	55/52/62	87/82/80	340/352/353	90/85/83
0.0075	60/59/68	90/87/85	465/464/471	95/88/87
0.01	68/68/76	92/90/88	**578/573/581**	**96/91/90**
0.0125	76/75/85	92/91/90	705/703/710	96/93/92
0.015	90/88/96	92/92/92	834/827/834	96/94/93
0.0175	**93/93/102**	**92/94/93**	961/949/958	97/94/93

AI (Nickl, 2023). The architect navigates this global quasi-territory: finding, modeling, and joining elements, and realizing them locally as multimodal assemblages.

CLIPedia thus becomes an architectonic instrument for assembling what we claim to know about the world we seek to address—and an architectural instrument for exploring the high-dimensional spaces that encompass much of what we can place in three dimensions.

5. Conclusion

CLIPedia provides a scalable, lightweight SOM-based architecture for local and notebook-based vector search, bespoke query and quest functions for navigating high-dimensional embedding spaces, and an open-source, reproducible setup adaptable to many models and corpora. Beyond these technical contributions, it offers a theoretical frame for reimagining the encyclopedia beyond its traditional structure. Theory informs not only the conceptual metaphors but also the computational methods, framing the SOM-based topology as an exploratory, interpretable environment distinct from conventional search engines or commercial LLM chatbot interfaces. By taking geodetic figures as metaphors and implementing them computationally, we chart paths across CLIPedia's geography — between points and poles, along geometric and probabilistic meridians. Stochastic scouts help us find islands in the vectorial archipelago; spectral compasses chart routes through digital forests; trained guides lead us to places we could never prompt or query. These are just some of the architectonic instruments architects might invent—to become not only users, but explorers and builders of the encyclopedias to come.

AI ethics statement

During the preparation of this work, the author used ChatGPT Plus during the coding and writing stages, reviewing and editing the content as needed, and takes full responsibility for the published article.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I thank my PhD supervisors, Prof. Ludger Hovestadt and Prof. Vera Bühlmann, for their support and for helping me build literacy in coding and theory. I am grateful to the Chair of Digital Architectonics at ETH Zurich and ATTP at TU Vienna—especially Julian Besems, Adil Bokhari, Dr. Cai Chenyi, Maya Christodoulaki, Dr. Pierre Cutellic, Dr. Jorge Orozco, Sophie Ramm, Dr. Miro Roman, Lorenzo Vicari, Dr. Riccardo Villa, Mo Yichen, and Dr. Elias Zafiris—for fostering an environment where this research could unfold, and to ITA for my Doctoral Fellowship. I also thank Prof. Li Biao and Prof. Tang Peng at SEU for inviting me to Nanjing, where CLIPedia was first presented at a conference. Lastly, I am grateful to Dr. Johanna Just for her editorial advice, to the editor for overseeing the publication process, to the anonymous reviewers for their helpful feedback, and to my past teachers and tutors for their guidance.

References

Aarseth, E., 2012. A narrative theory of games. In: Proceedings of the International Conference on the Foundations of Digital Games. Presented at the FDG'12: International Conference on the Foundations of Digital Games, ACM. Raleigh North Carolina, pp. 129—133.

Akyürek, E., Bolukbasi, T., Liu, F., Xiong, B., Tenney, I., Andreas, J., Guu, K., 2022. Towards tracing factual knowledge in Language Models back to the training data. In: Findings of the Association for Computational Linguistics. Presented at the EMNLP 2022. Association for Computational Linguistics.

Alberti, L.B., 2010. Ex ludis rerum mathematicarum. In: Williams, K., March, L., Wassell, S.R. (Eds.), The Mathematical Works of Leon Battista Alberti. Springer, Basel.

Alvarez Marin, D., 2020. Atlas of indexical cities: articulating personal city models on generic infrastructural ground. ETH Zurich.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. J. Mach. Learn. Res. 1137—1155.

Berners-Lee, T., 1989. The Original Proposal of the WWW, HTMLized.

BLIP-2. Hugging Face. Transformers documentation.

Bokhari, A., Nickl, A., 2024. Automata at court: characters, intelligences, machines. In: Kretzer, M. (Ed.), Synthetic Realities: New Frontiers in AI-Driven Design, Fabrication and Materiality (AADR).

Bratton, B.H., 2015. The stack: on software and sovereignty. In: Software Studies. MIT Press, Cambridge, Massachusetts.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language Models are Few-shot Learners.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y., 2023. Sparks of Artificial General Intelligence: Early Experiments With GPT-4 arXiv.2303.12712.

Bühlmann, V., 2024. The digital, a continent? Nature and poetics. In: Applied Virtuality Book Series. Birkhäuser. De Gruyter, Basel.

Bühlmann, V., 2020. Mathematics and Information in the Philosophy of Michel Serres, first ed. Bloomsbury Publishing Plc.

Cai, C., Wang, X., Li, B., Herthogs, P., 2024. ArchiSearch: A Text-Image Integrated Case-based Search Engine in Architecture Using Self-organizing Maps. Presented at the CAADRIA 2024: Accelerated Design, Singapore, pp. 19—28. https://doi.org/10.52842/conf.caadria.2024.3.019.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J., 2023. Reproducible scaling laws for contrastive language-image learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818—2829.

CLIP Multi-domain Feature Extractor. Wolfram Neural Net Repository.

Common Crawl, n.d. Open repository of web Crawl data.

de Kunder, M.. The size of the world wide web (the internet) [WWW Document]. WorldWideWebSize.com.

Debord, G., 2007. Theory of the derive. In: Knabb, K. (Ed.), Situationist International Anthology. Bureau of Public Secrets, Berkeley.
Eco, U., 2009. The Infinity of Lists. Rizzoli, New York.
Ethz, C.A.A.D., Saldaña Ochoa, K., Zifeng, G., 2020. Map & Models - Self Organizing Maps.
Forman, D., 2018. Leibniz on human finitude, progress, and eternal recurrence: the argument of the 'Apokatastasis' essay drafts and related texts. Oxf. Stud. Early Mod. Philos. VIII.
France, S.L., Douglas Carroll, J., Xiong, H., 2012. Distance metrics for high dimensional nearest neighborhood recovery: compression and normalization. Inf. Sci. 184, 92—110.
Gödel, K., 1931. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monat. Mathemat. Phys. 173—198.
Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning, Adaptive Computation and Machine Learning. The MIT Press, Cambridge, Massachusetts.
Heidegger, M., 2002. Off the Beaten Track. Cambridge University Press, Cambridge ; New York.
Hodel, T., 2024. Large Language Models, oder weshalb wir künstliche Intelligenz im Archiv finden sollten. In: Fähle, D., Müller, P. (Eds.), Smart und intelligent - Digitale Unterstützung für die Arbeit im Archiv, Werkhefte des Landesarchivs Baden-Württemberg. Wuerttembergische Landesbibliothek, Stuttgart.
Hovestadt, L., 2020. Writing & code. In: Hovestadt, L., Hirschberg, U., Fritz, O. (Eds.), Atlas of Digital Architecture: Terminology, Concepts, Methods, Tools, Examples, Phenomena. De Gruyter, pp. 369—397.
Hovestadt, L., 2015. Elements of a digital architecture. In: Bühlmann, V., Hovestadt, L. (Eds.), Coding as Literacy: Metalithikum IV, Applied Virtuality Book Series. Birkhäuser, Basel.
Jemielniak, D., 2019. Wikipedia: why is the common knowledge resource still neglected by academics? GigaScience 8, giz139.
Jina, A.I., 2024. jina-clip-v2 [WWW Document].
Jurafsky, D., Martin, J.H., 2025. Speech and language processing: an introduction to natural language processing. In: Computational Linguistics, and Speech Recognition with Language Models, third ed.
Kohonen, T., 2001a. Self-organizing maps. In: Springer Series in Information Sciences, 30th ed. Springer, Berlin Heidelberg, Berlin, Heidelberg.
Kohonen, T., 2001b. Self-organizing maps. In: Springer Series in Information Sciences. Springer, Berlin Heidelberg, Berlin, Heidelberg.
Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59—69.
Koukounas, A., Mastrapas, G., Günther, M., Wang, B., Martens, S., Mohr, I., Sturua, S., Akram, M.K., Martínez, J.F., Ognawala, S., Guzman, S., Werk, M., Wang, N., Xiao, H., 2024. Jina CLIP: Your CLIP Model is Also Your Text Retriever. arXiv.2405.20204.
LAION. Large-scale artificial intelligence open network [WWW Document].
Leibniz, G.W., 2017. The Horizon of Everything Human.
Lambert, J.H., 1965. Anlage zur Architectonic, oder Theorie des Einfachen und des Ersten in der philosophischen und mathematischen Erkenntniß, Philosophische Schriften. Georg Ohlms Verlagbuchhandlung, Hildesheim.
Leibniz, G.W., 1903. Opuscules et Fragments Inédits de Leibniz. Félix Alcan (Paris).
Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: Bootstrapping Language-image Pre-training With Frozen Image Encoders and Large Language Models. arXiv.2301.12597.
Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J., 2022. Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning.
Lin, S.-C., Lee, C., Shoeybi, M., Lin, J., Catanzaro, B., Ping, W., 2025. MM-embed: Universal Multimodal Retrieval With Multimodal LLMs. arXiv.2411.02571.
Lynch, K., 1960. The Image of the City. The MIT Press, Cambridge, Massachusetts; London, England.
Marincic, N., 2021. Numbasom.
Marinčić, N., 2019. Computational models in architecture: towards communication in CAAD: spectral characterisation and modelling with conjugate symbolic domains. In: Applied Virtuality Book Series. Birkhauser Verlag, Basel, Switzerland; Boston [MA].
Maynez, J., Narayan, S., Bohnet, B., McDonald, R., 2020. On faithfulness and factuality in abstractive summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. Online, pp. 1906—1919.
Mazanec, J., Hamzaoui, O., 2022. Choose the k-NN algorithm for your billion-scale use case with OpenSearch | AWS Big Data Blog. AWS Big Data Blog.
Microsoft, 2025. microsoft/DiskANN.
Models, 2025. Hugging Face.
Mount, N.J., Weaver, D., 2011. Self-organizing maps and boundary effects: quantifying the benefits of torus wrapping for mapping SOM trajectories. Pattern Anal. Appl. 14, 139—148.
Nickl, A., 2025. Clipedia.
Nickl, A., 2023. Matter mutter: talking things. Scroope Camb. Archit. J.
Nickl, A., Bokhari, A., 2023. LLMs in the city: contracting non-human characters. In: Markopoulou, A., Farinea, C., Marengo, M. (Eds.), Responsive Cities: Collective Intelligence Design Symposium Proceedings 2023. Presented at the Responsive Cities, IAAC. Barcelona.
OED. Encyclopaedia | encyclopedia, N., sense 1. [WWW Document]. Oxford English dictionary.
OpenCLIP Multi-domain Feature Extractor, 2023. Wolfram Neural Net Repository.
Orozco, J., 2017. Indexical Architecture: Prominent Positions, Applications and the Web. ETH Zurich.
Plant, S., 1997. Zeroes and Ones Digital Women and the New Technoculture. Fourth Estate, London.
Popper, K., 1998. The World of Parmenides. Routledge, London.
Qian, G., Sural, S., Gu, Y., Pramanik, S., 2004. Similarity between Euclidean and cosine angle distance for nearest neighbor queries. In: Proceedings of the 2004 ACM Symposium on Applied Computing. Presented at the SAC04: the 2004 ACM Symposium on Applied Computing. ACM, Nicosia Cyprus, pp. 1232—1237.
Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning Transferable Visual Models from Natural Language Supervision.
Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical Text-conditional Image Generation With CLIP Latents.
Roman, M., 2021. Play Among Books: A Symposium on Architecture and Information Spelt in Atom-Letters. De Gruyter.
Saldaña Ochoa, K., 2021. Event protocol: enhancing disaster response with architectonic capabilities by leveraging machine and human intelligence interplay. ETH Zurich.
Sammut, C., Webb, G.I. (Eds.), 2017. Encyclopedia of Machine Learning and Data Mining. Springer, US, Boston, MA.
Selcer, D., 2007. The uninterrupted ocean: Leibniz and the encyclopedic imagination. Representations 98, 25—50.
Serres, M., 2022. Hermes 2 Interference.
Serres, M., 2021. Hermès I: Communication. Randolph Burks. Éditions de Minuit, Paris, 1969.
Serres, M., 2018. The Incandescent. Bloomsbury Publishing PLC, London.

Serres, M., 2015. Thumbelina: the Culture and Technology of Millennials. Rowman & Littlefield International, London ; New York.
Serres, M., 1982. Le système de Leibniz et ses modèles mathématiques: Etoiles - Schémas - Points, second ed. Épiméthée. Pr. Univ. de France, Paris.
Simondon, G., 2017. On the Mode of Existence of Technical Objects. Univocal Publishing, Minneapolis, MN.
Subramanya, S.J., Kadekodi, R., Devvrit, Kadekodi, R., Krishaswamy, R., Simhadri, H.V., 2019. DiskANN: fast accurate billion-point nearest neighbor search on a single node. In: NeurIPS 2019. Presented at the 33rd Conference on Neural Information Processing Systems (Vancouver, Canada).
Touvron, H., Martin, L., Stone, K., 2023. Llama 2: Open Foundation and Fine-tuned Chat Models (GenAI).
Upreti, N., Codella, J., Simhadri, H., 2024. Azure Cosmos DB Vector Search with DiskANN Part 1: Full Space Search. Azure Cosmos DB Blog.
Venturi, R., Scott Brown, D., Izenour, S., 1988. Learning from Las Vegas: the Forgotten Symbolism of Architectural Form. MIT Pr, Cambridge, Mass.
Virilio, P., 2005. The Information Bomb, Radical Thinkers. Verso, London; New York.
Vitruvius, 2006. Ten Books on Architecture. Project Gutenberg, Salt Lake City.
Wales, J., 2004. Wikipedia founder Jimmy Wales responds - slashdot.
Wikimedia Commons [WWW Document]. Wikimedia foundation.
Wikimedia Foundation. Index of/enwiki.
Witt, A., 2022. Formulations: Architecture, Mathematics, Culture, Writing Architecture Series. The MIT Press, Cambridge, Massachusetts.
Wittgenstein, L., 1976. Wittgenstein's Lectures on the foundations of mathematics. In: Cambridge, 1939: from the notes of R. G. Bosanquet, Norman Malcolm, Rush Rhees and Yorick Smythies. Harvester Press, Hassocks.
Wolfram. The background and vision of mathematica—Stephen Wolfram writings.
Zaghloul, M., 2017. Machine-Learning Aided Architectural Design - Synthesize Fast CFD by Machine-Learning. ETH Zurich.
Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, Meishan, Li, W., Zhang, Min, 2025. GME: improving universal multimodal retrieval by multimodal LLMs.
Zifeng, G., 2021. From simulation to synthesis: architectural modeling with context-based encoding using data-driven computational machines. ETH Zurich.