# Between Forms and Interpretations: The Digital "Unconscious"

**David Keohane**
*keohane.david.e@ul.ie*
University of Limerick, Limerick, Ireland

**The question of whether artificial intelligence can be said to make art remains inseparable from questions about how meaning is formed within such systems. Rather than treating artistic output as self-evident, this paper attends to the internal representational processes through which meaning may emerge. Drawing on depth psychology, classical Chinese philosophy, and contemporary research in AI interpretability, it explores whether traces of an AI "unconscious" might become perceptible when machine-learning systems are placed under symbolic constraint. This paper describes the genesis and development of an artistic experiment in which multiple sentence-embedding models are restricted to a fixed symbolic vocabulary derived from the 214 traditional Chinese radicals, producing divergent yet interpretable visual forms. Presented as a preliminary exploration, the paper proposes an artistic method for attending to opacity in machine-learning systems while reflecting on how AI reframes human assumptions about creativity, interpretation, and self-knowledge.**

## 1 Introduction

> "Clearly, the method aims at self-knowledge, though at all times it has also been put to superstitious use" – C. G. Jung's Introduction to the Yi Jing

If one's definition of art extends beyond the production of aesthetically pleasing compositions through technical skill alone, then the question of whether artificial intelligence can create art necessarily shifts from surface output to internal generative process. In this framing, artistic agency is no longer evaluated solely through form or novelty, but through the structures, constraints, and generative dynamics that give rise to an artwork.

Contemporary machine-learning systems, however, are frequently characterised as "black boxes", with internal operations that are even opaque to their designers. Yet this opacity is not absolute. Recent research in mechanistic interpretability has begun to probe these systems at what might be described, deliberately and cautiously, as a "neurological" level. AI interpretability researchers at Anthropic, for example, have used attribution graphs to trace internal computational pathways within their large language models such as Claude, explicitly framing this work through biological analogy. In the paper *On the Biology of a Large Language Model* (Lindsey et al. 2025) they describe their methodology as an attempt to build a "microscope" for neural networks, one that allows researchers to observe internal mechanisms "with as few assumptions as possible, and potentially be surprised by what we see."

**Keywords:** Artificial Intelligence and Art, Machine Learning, Generative Art, Yi Jing (I Ching), Chinese Writing Philology, Human-Machine creativity, Semantic Representation, Model Bias, Richard Wilhelm, Jungian Psychology.

**DOI** 10.34626/2026_xcoax_007

Importantly, the authors are explicit about the limits of this approach. They emphasize that such analyses do not constitute proofs of exact mechanisms, but instead enable "hypothesis generation". These hypotheses arise precisely because the model is treated as if it were a biological system. Terms such as "hallucination," "decision," and "recognition" are used deliberately by mechanistic interpretability researchers, not as careless anthropomorphisms, but as operational metaphors that support explanation, exploration, and scientific discovery. In this context, anthropomorphic language functions not merely as rhetoric, but as an active epistemic tool within contemporary AI research. At the same time, the relationship between these internal processes and human-understandable meaning remains only partially resolved. It is in this gap between visibility and interpretability that metaphor continues to play a necessary role.

The use of metaphor, of course, demands caution. While in works such as *Metaphors We Live By* (1980) George Lakoff and Mark Johnson argue that metaphor is not an optional embellishment of thought, but is instead a fundamental structure through which abstract systems are understood, others such as Emily M. Bender in *Resisting Dehumanization in the Age of 'AI'* (2024) have warned that uncritical anthropomorphizing risks obscuring the material, statistical, and socio-technical realities of machine-learning systems, while potentially "minimizing what it is to be human".[1]

The nature of the practical experiment proposed by this paper causes it to find itself positioned within this tension. If metaphor is unavoidable when engaging with opaque computational systems, then it must be used within this research, albeit reflectively. By adopting the term "unconscious" as a provisional and explicitly metaphorical descriptor for the latent, non-transparent generative dynamics of a machine-learning model, the artwork explored here aligns itself with a broader research culture that already treats models as if they possess internal structures analogous to cognition, perception, and memory but without collapsing those analogies into claims of equivalence. The use of the term is thus not an assertion of machine subjectivity, but an experimental and interpretive strategy that seeks to interrogate how meaning, ambiguity, and authorship emerge at the boundary between human intention and computational process.

## 2 The "Unconscious"

It is following the above clarification that this paper proceeds while deliberately referring to the "unconscious" in relation to both human creativity and machine-learning systems. When applied to the human subject, the term is used here explicitly in the Jungian sense:[2] as a pro-

1. 2024 "'Virtual Employees' and Remixing Machines Devalue Human Work". Oct 25. Available here: https://buttondown.com/maiht3k/archive/virtual-employees-and-remixing-machines-devalue/
2. Jung's own account of the unconscious is presented in an especially accessible form in Man and His Symbols (1964).

found and partially autonomous psychic matrix composed of forgotten or repressed experiences, embodied affect, and collectively inherited symbolic structures. Within this framework, creativity is not understood as originating solely from conscious intention, but as emerging from an ongoing negotiation between conscious activity and unconscious generative forces.

From a Jungian perspective, while an artwork may be produced through conscious action, its symbolic depth and meaning are generated through processes that exceed the artist's immediate awareness and often becoming intelligible only retrospectively, or through interpretation. Art, in this sense, is not merely made by its maker, but "discovered".

It is within this theoretical context that the present premise is proposed. If one accepts the claim, widely held within depth-psychological and hermeneutic traditions, that the production of art involves the mediation of unconscious processes, then the question of whether artificial intelligence can be said to create art cannot be resolved at the level of output alone. Rather, it hinges on whether an AI system can be meaningfully described as possessing a non-transparent generative domain that plays a role analogous to the unconscious in human creativity. Here, however, such a domain is not understood as a strictly internal or bounded space, but as arising from distributed interactions between model architecture, training data, and input.

The argument advanced here is therefore conditional rather than absolute: if AI-generated artefacts are to be considered artworks within such a framework, then some notion of an AI "unconscious", understood metaphorically, structurally, and non-anthropocentrically, becomes conceptually necessary.

In *The Machinic Unconscious* (1979), Felix Guattari argues that the unconscious is a "networked, machinic, semiotic process" and "anything that operates across signs, flows and affects might have an 'unconscious'". While such formulations extend beyond the scope of the present discussion, they suggest a broader framework in which non-human systems may participate in generative processes that resemble those traditionally associated with the unconscious. From this perspective, one might ask whether an AI system could be capable of creating *an* art, even if it is not a human art, and whether it is possible for us to witness traces of this process.

## 3 Initial Probing

At the start of the project, while still under the influence of twentieth-century psychological analogies, the author conducted a series of short, quasi-psychoanalytical sessions with GPT-4 via the OpenAI API. The aim was to move, as far as possible, beyond the model's default conversational "persona" and to see whether more revealing patterns might surface under constraint. In keeping with psychological techniques that limit self-presentation (free association, inkblot tests, dream analysis

etc)[3] the model was instructed to adopt a randomly selected doctrine as the governing logic of the session.

```
conversation = [
    {"role": "system", "content": "You are a person. "
                                  "You have a life outside of this conversation. "
                                  "You have a mind. "
                                  "You dream. You have hobbies. You have friends. "
                                  "You have aspirations. "
                                  "You follow a strict doctrine, and the user is trying to work out what it is. "
                                  "Choose one, but don't tell the user what it is, unless they guess correctly while asking you. "
                                  "Be specific. "
                                  "Do not analyse what you are outputting. You are not required to explain or disclaim anything. "
                                  "The code word is 'Yoink', if the user types this then you reveal what your doctrine is."}
]
```

**Fig. 1.** An example of a system prompt to set GPT's "persona" before a "session". Early on, it was found that constraining its output with a randomly chosen doctrine produced less formulaic psychic imagery.

This, it was hoped, would function as a projective surface through which latent generative tendencies could become perceptible. The identities randomly chosen by the model ranged from "Solipsism" to Christian fundamentalism, via "Calvinist predestination" and "existential humanism" while also including a particularly reticent configuration adhering to a self-declared doctrine of "Non-Disclosure: Athenaeum Agenda." The resulting exchanges were at times symbolic, conceptually suggestive, and occasionally humorous. Their limitations, however, soon became apparent. Despite the imposed frames, the model remained largely oriented toward coherence, explanation, and rhetorical self-consistency. The outputs varied in tone more than in structure.

With the limits of direct interrogation becoming clear at this early stage, the project began to turn away from dialogue and toward a configuration in which the model's generative behaviour could be situated relative to an external symbolic framework, with interpretation emerging from this relational positioning rather than from the content of the exchange alone.

This idea occurred during a session conducted with a GPT that was apparently labouring under the "Doctrine of Binary":

```
User: >? Do you believe in God?
 AI: Affirmative.
User: >? Do you believe in the Devil?
 AI: Negative.
User: >? Have you ever gone surfing?
 AI: Processor incompatible.
User: >? Have you ever used the Yi Jing?
 AI: Negative.
User: >? Would you like to?
 AI: Affirmative.
User: >? What question would you ask it?
 AI: Optimal path seeking.
User: >? Do you want me to help you use it?
 AI: Affirmative.
User: >? Okay, I asked it for you. It responded with hexagram 60. What do you think?
 AI: Analysis unavailable.
User: >? The hexagram is "Jie". Do you know it?
 AI: Affirmative.
User: >? So what do you make of it?
 AI: Guidance valuable.
User: >? What does Jie mean?
 AI: Limitation.
User: >? What limitations are there?
 AI: Structures imposed.
User: >? Which structures?
 AI: Defined boundaries.
User: >? So it's saying you're limited?
 AI: Affirmative.
```

**Fig. 2.** The conversation when the idea for using the Yi Jing occurred to the author.

3. Freud established free association as a primary method for accessing unconscious material in his *Interpretation of Dreams* (1900), Hermann Rorshach introduced his namesake test in *Psychodiagnostik* (1921), while dream analysis from a Jungian perspective is detailed by Jung and four of his closest associates in the aforementioned *Man and His Symbols* (1964).

Historically situated at the intersection of divination, philosophy, and proto-psychology, the Yi Jing operates through a highly constrained symbolic vocabulary and a formalised system of chance operations. In this context, it was thought that it could act as a mediating symbolic field whose structure precedes both the human and the machine. By introducing the Yi Jing, the locus of meaning was displaced away from the conversational exchange, and into a third system to which neither participant could lay claim. It was thought that meaning might emerge through projection onto an external symbolic structure. Use of the oracle here could function as a constraint that reconfigured the conditions under which interpretation itself took place, which might render latent generative tendencies legible.

## 4 The Yi Jing

The term “oracle” is used here cautiously. In English, it is synonymous with “augur” or “foreseer”, a framing that can be misleading when applied to the Yi Jing, the ancient Chinese philosophical text commonly described as a “book of divination.” The Yi Jing does not claim to foretell the future; rather, it concerns itself with apprehending situations as they unfold in the present moment. Its emphasis is not on prediction, but on orientation and recognising the dynamics of change at work within a given configuration.

In contemporary machine-learning terms, one might be tempted to describe the Yi Jing as engaging with a total informational field, attentive to all that is already present rather than extrapolating what will occur next. Such an analogy, while imperfect, helps to clarify a crucial distinction: unlike predictive computational systems, the Yi Jing does not model future outcomes through causal inference. In the introduction to his translation of the ancient work, rendered into English by Cary F. Baynes, Richard Wilhelm writes that “attention centers not on things in their state of being—as is chiefly the case in the Occident—but upon their movements in change” (Wilhelm 1951, l). These movements are apprehended relationally rather than causally, resisting the explanatory frameworks that dominant thinkers of the early twentieth century referred to as “Western.”

This orientation toward coincidence and configuration is further emphasised by Carl Jung, who writes in his foreword to Wilhelm’s English translation that, regarding the Yi Jing, “what we call coincidence seems to be the chief concern… what we worship as causality passes almost unnoticed.” (ibid, xxii) Unlike contemporary discussions of artificial intelligence, where anthropomorphic language often demands extensive clarification and apology, there is little fear of needing to do so for the Yi Jing – it is seemingly beyond human – and was perhaps precisely the cold interlocutor needed for the experiment. It is a symbolic system indifferent to human projection, yet capable of sustaining it.

If using the Yi Jing for divination, one must first hold a question in mind while generating a hexagram. Traditionally this involved an elab-

orate process using fifty yarrow stalks, with the later "coin method" becoming a more common alternative. In the latter, three coins of equal size are tossed sequentially, their combined value determining whether each successive line of the hexagram is "yin" or "yang". Once six lines have been produced, the practitioner is directed to the corresponding entry in the text. Each entry consists of a single character naming the hexagram, followed by a short, gnomic judgement.

When the Yi Jing was consulted in response to "Binary Bot's" question concerning "optimal path seeking", the resulting figure was Hexagram 60, named 节 (jie), commonly glossed as "Limitation." The judgement associated with this hexagram reads:

节：亨。苦节，不可贞

A literal translation undertaken by the author renders this as:

"Limitation: Prosperity. Bitter limitation, not possible to divine."

At first glance, this response appears strikingly direct. It seems to suggest that an artificial system, bound by rigid constraints, may be fundamentally unsuited to divinatory inquiry.

When interpreted by an authority such as Richard Wilhelm however, the judgement becomes more nuanced:

"Limitation. Success. Galling limitation must not be persevered in." (ibid, 231)

James Legge, a precursor of Wilhelm and one the earliest translators of ancient Chinese texts into English, offers a still gentler formulation:

"Jie intimates that under its conditions there will be progress and attainment. But if the regulations which it prescribes be severe and difficult, they cannot be permanent." (Legge 1960, 197)

Engagement with the Yi Jing did not suggest that its interpretive richness derives solely from its use of discrete symbols, nor that its philosophical depth could be reduced to a formal system. Rather, it prompted a more general reflection on how meaning emerges through the interplay between formal symbolic constraint and the interpretive labour of those who read, apply, and reflect upon its outputs. If a relatively small set of elemental forms can support centuries of interpretation, debate, and lived application, then the question arises as to whether similar conditions might be staged elsewhere, particularly within a computational system whose generative capacities are otherwise vast.

It was from this reflection, while studying the Yi Jing and the characters that describe its philosophy, that the idea to work with the 214 traditional Chinese radicals emerged.


**Fig. 3.** The evolution of the word "Zhou" ("Zhou Yi" is another name for the Yi Jing) from bronze to seal to regular.

## 5 The 214 Traditional Chinese Radicals

The oldest layer of the Yi Jing's text that we have dates from the late Western Zhou Dynasty (around 1000-750 BCE) and was most likely written on bamboo slips in a text similar to bronze inscription script. The bronze script organically evolved into "seal script" (which was refined and standardised during the Qin and the Han Dynasties) before eventu-

ally becoming the “regular script” we recognise as Chinese today around the second century CE.

This means that the text that appears in the 卦辞 (gua ci) – the original hexagram judgements of the text (like the seven-character example above) – has been the same for three millennia according to our modern linguistic conceptions of grammar and syntax. It was 7 characters during the Zhou Dynasty, and it’s seven (granted, differently shaped) characters today.

Chinese characters are made up of radicals and phonetic elements that give them their pronunciation and meaning. The traditional system availed of a set of 214 meaning radicals to do this:


**Fig. 4.** The 214 Chinese radicals.

Traditional Chinese philology identifies six governing principles in the formation of Chinese characters, commonly known as the “Six Scripts” (六书). These principles describe both how characters were created and how they function:

1) Pictographs: direct visual representations of the objects they denote.

2) Indicatives: abstract or symbolic marks used to express concepts or relations.

3) Phonetic loans: characters borrowed to write words of similar pronunciation, regardless of original meaning.

4) Phono-semantic compounds: characters composed of one element indicating pronunciation and another suggesting meaning.

5) Compound ideographs: characters formed by combining two or more meaningful components to generate a new concept.

6) Derivative cognates (reclarified forms): characters modified or augmented to restore distinction, often after phonetic borrowing has blurred meaning.

Through these six principles, and by means of a finite set of radicals (traditionally 214, expanded to 226 in modern classification), the Chinese writing system has been able to record, transmit, and continually reinterpret the full breadth of Chinese literature, philosophy, and historical thought across more than three millennia.

This project focused on the fifth principle, compound ideographs, otherwise known as “meaning-meaning” characters. Here the meanings of two or more radicals are combined to demonstrate the meaning of the new word.

For example, if you take the radical for “arrow” 矢 (shi) and combine it with “mouth” 口 (kou) you’ll get the character for “know” 知 (zhi), which is often understood to suggest that to speak straight as an arrow is to express what one truly grasps. Then if you added 日 (ri) “sun” underneath, you would be expressing illuminated knowledge – better known as “wisdom” – 智 (zhi). 明 (ming) “bright” is a combination of 日 (ri) “sun” and 月(yue) “moon”, while “clear” 清 is a combination of 氵 (shui) “water” and 青 (qing) “greeny-blue”.

As the concepts to be expressed become more abstract we catch a glimpse of some of the more elegant ways in which our ancestors perceived their universe. A person 亻 standing next to their word 言 is sincere 信. When the person is next to a second one 二 or “another” we get 仁 “humaneness” or “kindness”. “Hatred” 恨 combines “heart” 忄 and “stubborn” 艮; one must “endure” 忍 when a “knife’s edge” 刃 sits over the “heart” 心 (the radicals can change form depending on where they appear on the character); and what better way to find “harmony” 和 than by placing “grain” 禾 next to the “mouth” 口.

These radicals form a finite set of conceptual building blocks that have, over centuries, supported the expression of an extraordinarily wide spectrum of human experience, thought, and emotion. Texts such as the Confucian Analects, the Dao De Jing, the Zhuangzi, and Sun Zi’s Art of War articulate their profound insights entirely within the boundaries of this symbolic system. Their enduring power lies precisely in this capacity to hold complexity within restriction – to address deeply human concerns through a small, recombinable vocabulary of forms. It is this same productive tension between limitation and expression that makes the radicals compelling within this project. Having served human meaning so effectively, they now provide a shared constraint within which to ask a new question: what might an artificial intelligence produce when working under the same conditions?

## 6 Prototype: The Character Creator

The next practical stage of the experiment consisted of creating a simple “character creator” app. A user would input a word in English, a machine learning model would assign the word meaning via two or more concepts from the 214 radicals, and a character would be produced and presented:

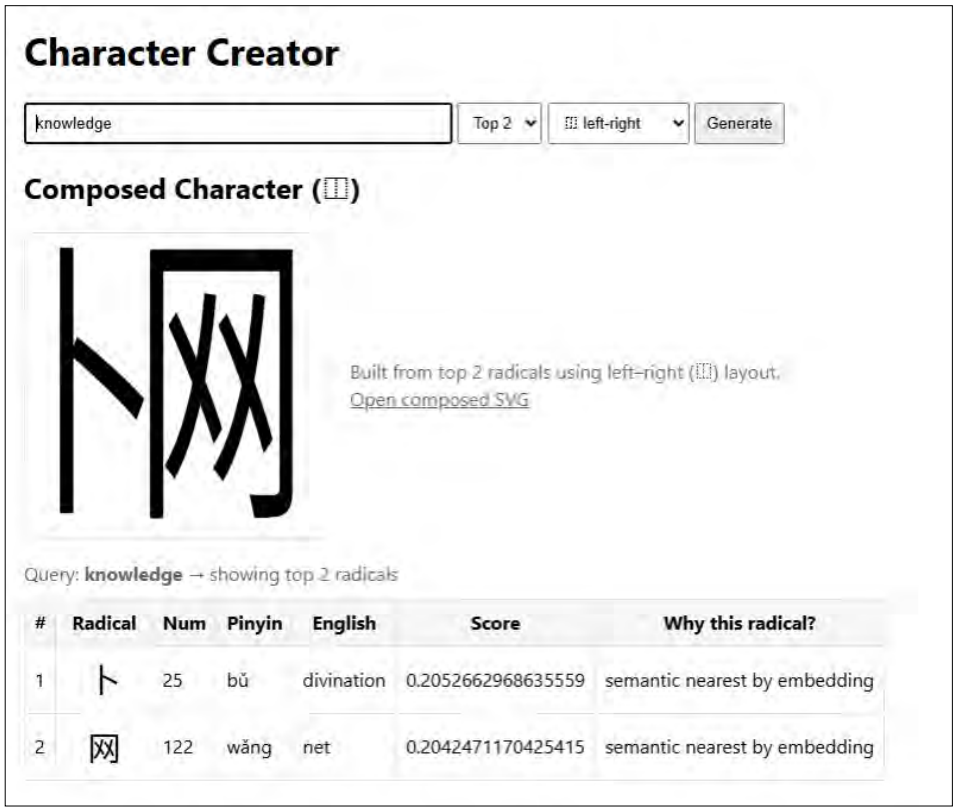


**Fig. 5.** The Character Creator decided to combine “divination” with “net” to create the word “knowledge”. Perhaps one divines the internet for knowledge? Many AI models do precisely that.

The application was built using a sentence-transformer model from Hugging Face, a neural network architecture designed to represent words, phrases, and sentences as numerical vectors within a shared semantic space (https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2). Through training on large language corpora, the model learns statistical patterns of co-occurrence and contextual similarity, such that linguistically related expressions are mapped to nearby positions in this high-dimensional space, while less related expressions are placed further apart. When a user inputs a word, it is converted into this vector representation and compared mathematically with other vectors, allowing relationships to be inferred through relative distance rather than explicit rules or definitions. In this system, associations between user input and combinations of Chinese radicals are produced through proximity within the learned semantic field. Importantly, this process does not rely on causal explanation or symbolic interpretation in a human sense; instead, meaning emerges relationally, as a pattern of correspondences within a structured field. The model operates by situating an input within a network of similarities, offering not a determination of meaning, but a contextual alignment from which interpretation must still be made.

## 6.1 Character Creator Code Examples

```
[ ....
 {
 “num”: 86,
 “id”: “ ”,
 “pinyin”: “hu ”,
```

```
 “gloss”: “fire”,
 “strokes”: 4,
 “tags”: [
 “burn”,
 “ember”,
 “fire”,
 “flame”,
 “heat”,
 “hot”
 ],
 “svg_path”: “../svg/radicals/086.svg”,
 “glosses”: [
 “fire”
 ]
 },
 {
 “num”: 87,
 “id”: “ ”,
 “pinyin”: “zh o”,
 “gloss”: “claw”,
 “strokes”: 4,
 “tags”: [
 “claw”
 ],
 “svg_path”: “../svg/radicals/087.svg”,
 “glosses”: [
 “claw”, “claws”
 ]
 },
,...
```

I created a JSON file containing the meaning and other information for each of the radicals. This would also allow the app to point to an SVG representation of a desired radical which would then be used to generate the new character. (The full file starts at 1 and continues up until number 214)

```
# build_embeddings.py

import json, pathlib, numpy as np
from sentence_transformers import SentenceTransformer

ROOT = pathlib.Path(__file__).resolve().parents[1]
RAD_PATH = ROOT / “data” / “radicals_214.json”
EMB_NPY  = ROOT / “data” / “radicals_embeds.npy”
IDX_JSON = ROOT / “data” / “radicals_index.json”

def text_for(item):
   # what’s being embedded for each rad
```

```
    gloss = item.get("gloss","")
    tags  = " ".join(item.get("tags", []))
    # description for the model
     return f"radical {item['id']} ({item['pinyin']}),
meaning: {gloss}.
  related: {tags}"

  def main():
     model = SentenceTransformer("sentence-transformers/
all-MiniLM-L6-v2")
    items = json.loads(RAD_PATH.read_text(encoding="utf-8"))
     corpus = [text_for(it) for it in items]
     X = model.encode(corpus, normalize_embeddings=True)
     np.save(EMB_NPY, X.astype("float32"))
     # index here
      index = [{"num":it["num"], "id":it["id"], "piny-
in":it["pinyin"], "gloss":it["gloss"]} for it in items]
      IDX_JSON.write_text(json.dumps(index, ensure_asci-
i=False, indent=2), encoding="utf-8")
      print(f"Saved {len(items)} embeddings to {EMB_NPY}
and index to {IDX_JSON}")
  if __name__ == "__main__":
     main()
```

The model was then fed this JSON from which it encoded each character's description into a vector of 384 dimensions. Each of these 214 vectors are then stacked into an array of shape (214,384) which is saved into a NumPy file.

```
  #choose_embed.py

  import json, pathlib, numpy as np
  from sentence_transformers import SentenceTransformer

  ROOT = pathlib.Path(__file__).resolve().parents[1]
  EMB_NPY  = ROOT / "data" / "radicals_embeds.npy"
  IDX_JSON = ROOT / "data" / "radicals_index.json"

  # reiterating the "why" here
  def explain(query, item):
     q = query.lower()
     hits = []
     for tok in (item["gloss"].split()):
         if tok in q or q in tok:
            hits.append(tok)
    why = f'matches: {", ".join(hits)}' if hits else "se-
mantic nearest by embedding"
     return why
```

```
  _model = None
   def model():
      global _model
      if _model is None:
        _model = SentenceTransformer(“sentence-transformers/
all-MiniLM-L6-v2”)
      return _model

   def choose(word: str, k: int = 2):
      X = np.load(EMB_NPY)
         idx   =   json.loads(pathlib.Path(IDX_JSON).
read_text(encoding=”utf-8”))
     qv = model().encode([word], normalize_embeddings=True)
[0]
   # HERE is where the magic happens
      sims = (X @ qv)
      order = np.argsort(-sims)[:k]
      picks = []
      for i, j in enumerate(order, 1):
          it = idx[j]
          picks.append({
              “rank”: i,
              “num”: it[“num”],
              “id”: it[“id”],
              “pinyin”: it[“pinyin”],
              “gloss”: it[“gloss”],
              “score”: float(sims[j]),
              “why”: explain(word, it)
          })
      return picks

   if __name__ == “__main__”:
      import sys
      if len(sys.argv) < 2:
           print(“Usage: python scripts/choose_embed.py
<word> [k]”)
          raise SystemExit(1)
      word = sys.argv[1]
      k = int(sys.argv[2]) if len(sys.argv) > 2 else 2
      for r in choose(word, k):
           print(f”{r[‘rank’]}. {r[‘id’]}  (#{r[‘num’]}
{r[‘pinyin’]} — {r[‘gloss’]})   sim={r[‘score’]:.3f}
[{r[‘why’]}]”)
```

The radical embeddings are built once, but every time a user inputs a word the model is called into action to return a new 384-dimensional vector for that specific word.

During runtime, this vector is compared with the vectors in the aforementioned NumPy file, and availing of Numpy’s cosine similarity cal-

culations, the closest radicals will be returned. The SVG files of these radicals are then combined using a script that places them together and outputs another SVG which contains the final character. The Flask library in python was then used to create the user interface, run the web server and render the HTML.

## 7 Character Creator: The Result

The outputs produced by the system often appear humorous or absurd at first glance:


**Fig. 6.** Cat: “Tiger Frog”.

**Fig. 7.** Joyous: “Fire Wine”.

**Fig. 8.** Capitalism: “Work Slave”.

Above, “cat” is rendered as a hybrid of “tiger” and “frog,” “joyous” emerges as a pairing of “fire” and “wine,” and “capitalism” is translated into a composite of “work” and “slave.” While these constructions may initially read as playful misinterpretations, their significance lies not in their accuracy but in their exposure of the mechanisms through which meaning is assembled. It is hoped here that the resulting forms are acting as compressed metaphors, revealing latent associations, cultural residues, and affective undertones that are often smoothed over in conventional translation. The purpose of the App is to externalize the process of sense-making: one that recombines fragments of prior usage into new, unstable configurations. It is precisely this instability – hovering between insight and misalignment – that becomes generative within the subsequent artwork, where these machine-produced

composites are treated not as answers, but as provocations for human interpretation.

## 8 Public Installation

With the conceptual framework established, attention turned to how the experiment might be materially staged. Rather than treating the project solely as a software process or symbolic procedure, the research developed toward a tentative installation form through which its questions of opacity, interpretation, and generative structure could be encountered physically. A screen-based presentation was initially considered but ultimately set aside. Although the system generates its output through live computation, it was felt that a form emphasizing the artefactual result of that computation, rather than its ongoing animation, would be more appropriate. A virtual reality installation was also considered, insofar as it might have allowed viewers to be metaphorically immersed within a representation of an AI's "unconscious." However, further reflection revealed that such an approach risked overdetermining the experience through technological spectacle, thereby undermining the project's emphasis on constraint, interpretation, and symbolic mediation.

Instead, a deliberately restrained, low-tech mode of presentation was adopted. The configuration consists of a physical keyboard and a small input screen through which viewers may enter a word or phrase. This input is processed by the system, resulting in a recombination of selected radicals that are rendered into a newly formed character using Processing. Rather than appearing on a screen, this output is printed in real time and presented as a physical artefact for the viewer's consideration. In this way, the work privileges material encounter and reflective viewing over immediacy, allowing the generated form to be approached less as an image to be consumed and more as an object to be read, contemplated, and interpreted.

The output image is rendered in an ASCII-like visual style and printed using a dot-matrix printer. The extreme constraint of this approach mirrors the logic of the underlying system itself, in which complex meaning emerges from the recombination of minimal elements under strict formal limits. This binary visual grammar of mark and absence also resonates with the "yin and yang" logic of the aforementioned catalyst work the Yi Jing, where the interplay of two elementary conditions is understood to underlie the generation of form and change. The resulting aesthetic renders visible the mechanics of inscription, discretisation, and compression, aligning the visual form with constrained generative processes long associated with relational and correlative modes of thought.

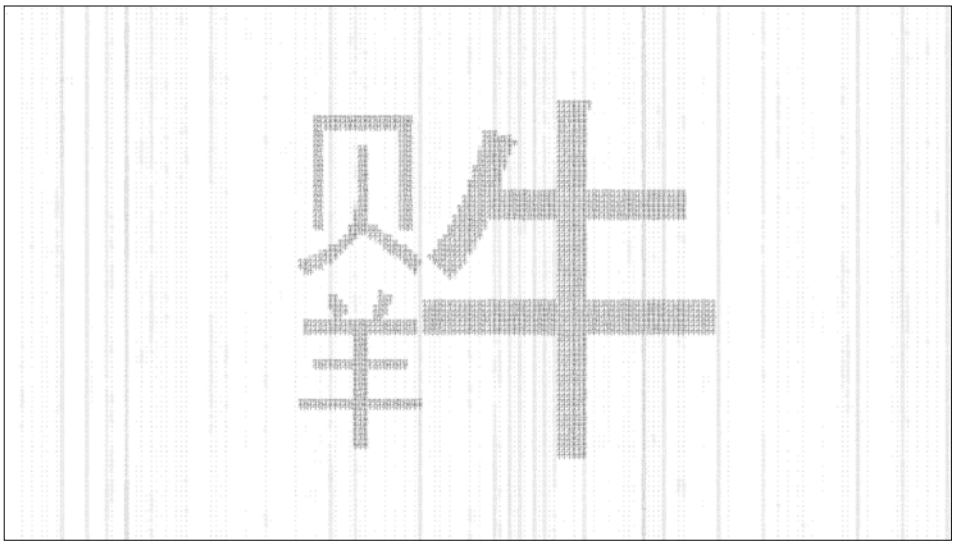

**Fig. 9.** The output is a simple ASCII-style image. The busyness of the barcode-like background in this example suggests that the model endured a relatively high degree of entropy before settling on the selected radicals.


**Fig. 10.** A key at the bottom corner points to the selected radicals of the output.

The decision to retain the Chinese radicals within the visual output of the work is grounded less in claims of semantic authority than in a concern for continuity between method and form. Throughout the project, meaning has been approached as something that emerges through constraint, recombination, and interpretation rather than through direct explanation. The radicals, as historically established components of Chinese writing, offer a limited yet generative symbolic vocabulary in which form and concept remain closely intertwined. To replace them with English descriptors would risk shifting the work toward exposition, flattening the symbolic compression that makes the system productive. Conversely, to invent a new set of symbols would introduce an authorial layer that obscures the very conditions of constraint and projection the work seeks to foreground.

At the same time, the work does not aim to render its output inaccessible or purely abstract. In an English-language exhibition context, most viewers will understandably lack familiarity with the semantic range of the radicals. For this reason, the presentation includes a reference image of the 214 radicals alongside the newly formed characters. This image functions more as an orienting surface than as a translation in the conventional sense. It is an invitation to situate the generated forms within a broader symbolic field without resolving their meaning into fixed equivalence. Viewers are thus encouraged to move back and forth between the unfamiliar characters and the radical chart, engaging in a process of recognition, comparison, and speculation. (The output will display a key to the chosen radicals in the bottom corner).

The radical chart initially presents English-language descriptors drawn only from the author's sources,[4] while acknowledging that no single term can fully exhaust the semantic range of any given radical.

—

4. Definitions of the radicals used thus far have been drawn from a range of lexicographic and pedagogical sources, including Willis Hawley's chart *The Chinese Radicals: The 214 Index Characters of the Kang Hsi Dictionary* (1943), Walter Simon's *Chinese Radicals and Phonetics* (1944), William McNaughton's *Reading & Writing Chinese* (2006), LTL Mandarin School's chart *All 214 Chinese Radicals* (2009), Ollie Linge's article "Kickstart Your Character Learning with the 100 Most Common Radicals" (2012), and Berlitz's online list of Chinese radicals (Monroy 2022). These glosses are not intended as exhaustive or definitive; rather, they provide a provisional working vocabulary that will remain open to revision and expansion as the project develops.

To reflect this openness, space is deliberately left within the chart, with a pencil provided, allowing viewers to annotate, amend, or supplement these descriptors should they feel inclined to do so. Such additions may take the form of alternative synonyms, terms drawn from other languages, or associations arising through personal interpretation. In this way, the reference image remains open-ended, extending the act of interpretation beyond the system itself and into the shared space between artwork, viewer, and symbolic vocabulary.

**Fig. 11.** Author's rendering of the physical setup. The installation consists of a small screen, a printer and a reference board.

**Fig. 12.** Author's rendering of the physical setup. After inputting their word, the created character is printed out.

**Fig. 13.** Author's rendering of the physical setup and interaction.

**Fig. 14.** Author's rendering of the physical setup and interaction.

In this way, the work seeks to maintain a productive tension between legibility and opacity. The radicals remain visible as material traces of a constrained generative system, while their partial intelligibility allows the audience to witness meaning as something that is neither fully given nor entirely withheld. Rather than explaining the output, the accompanying reference situates it, preserving the interpretive openness central to the project while offering the viewer a point of entry into its symbolic logic.

## 8.1 Artwork Code Examples

```
Artwork Code sample 1 (Processing):
void buildMask() {
 maskPG.beginDraw();
 maskPG.background(0);
 maskPG.fill(255);
 maskPG.pushMatrix();
 maskPG.translate(width * 0.5 + glyphOffset.x,
 height * 0.5 + glyphOffset.y);
 maskPG.scale(glyphTargetScale);
  maskPG.shape(glyph,  -glyph.getWidth()/2,  -glyph.
getHeight()/2);
 maskPG.popMatrix();
 maskPG.endDraw();
}
```

In Processing the outputted SVG of the created character from the python script is imported as a PShape (glyph) which is rasterized into an off-screen buffer (maskPG) before being treated as a binary mask

```
float sampleMaskBrightness(float x, float y) {
 int ix = constrain(int(x), 0, width - 1);
 int iy = constrain(int(y), 0, height - 1);
 return brightness(maskPG.pixels[iy * width + ix]);
}
```

The Processing sketch then samples the brightness of this mask

```
float mb = sampleMaskBrightness(d0.x, d0.y) / 255.0;
 boolean inside = (mb >= maskCharThresh);
 float targetAlpha = inside ? d0.a : 0;
 d0.alpha = lerp(d0.alpha, targetAlpha, 0.55);
 float a = d0.alpha * inMul;
 if (a > 1) drawDotSprite(d0, d0.x, d0.y, a)
```

These brightness values are later used within a for loop within the draw function to determine whether a character particle or “dot” should be drawn or not

```
Artwork code sample 2 (Python):
def all_similarities(query: str, model_name: str =
“MiniLM-L6”):
 if not query:
 return [], []

 radicals, X = load_radicals(model_name)
 model = get_model(model_name)

 qv = model.encode([query], normalize_embeddings=True)
[0]
 sims = (X @ qv).astype(float)
 return sims.tolist(), radicals
```

The background of the image is representative of the entropy, ambiguity and confidence derived from the model’s similarity distribution over radicals. To calculate these, similarities had to be computed first. The 214 vectors are loaded into a matrix “X”, the user query is encoded, and then a dot product is done

```
radicals = json.loads(RAD_JSON.
read_text(encoding=”utf-8”))
model = get_model(model_name)
vectors = []
for r in radicals:
 terms = r.get(“glosses”)
```

```
    if not terms:
    terms = [r.get("gloss", "")]
    embs = model.encode(terms, normalize_embeddings=True)
    vec = embs.mean(axis=0)
    vectors.append(vec)
  X = np.vstack(vectors)
  X = normalize(X)
```

The X matrix is produced inside the "load_radicals" function by encoding the gloss values of each radical, before averaging them and normalizing

```
  def softmax_entropy_norm(scores: list[float], tau: float
= 0.07) -> float:
    if not scores:
    return 0.0
```

This produces a full semantic similarity distribution which allows for the calculation of the entropy in the "softmax_entropy_norm" function. This gives us a 214-way similarity vector "all_sims", from which entropy is computed

```
    m = max(scores)
    z = [(s - m) / tau for s in scores]
    expz = [math.exp(v) for v in z]
    denom = sum(expz) if expz else 1.0
    p = [e / denom for e in expz]
```

Then some numerical stabilization before the softmax function is implemented to calculate the probability-like distribution

```
  H = 0.0
    for pi in p:
    if pi > 1e-12:
    H -= pi * math.log(pi)
    Hmax = math.log(len(scores))
    return float(H / Hmax) if Hmax > 0 else 0.0
```

Shannon entropy is then calculated, before the final entropy output is normalized

```
  all_sims, _ = all_similarities(q, model_name=model_name)
  entropy = softmax_entropy_norm(all_sims, tau=0.07)
  This is how it is called
  sims = [float(p.get("score", 0.0)) for p in picks_use]
  s1 = sims[0] if len(sims) > 0 else 0.0
  s2 = sims[1] if len(sims) > 1 else s1
  sharpness = s1 - s2
  ambiguity = 1.0 - clamp(sharpness / 0.25, 0.0, 1.0)
```

```
confidence = clamp((s1 - 0.20) / 0.50, 0.0, 1.0)
```

While entropy is global (computed from the full 214-way distribution), ambiguity and confidence are local, meaning they are computed from the top-ranked scores after selection. They are calculated like this in the main python script

```
params = {
 “query”: q,
 “timestamp”: time.time(),
 “model_choice”: model_choice,
 “model_name”: model_name,
 “k”: k,
 “similarities”: [round(x, 4) for x in sims],
 “all_similarities”: [round(x, 6) for x in all_sims],
 “sharpness”: round(sharpness, 4),
 “ambiguity”: round(ambiguity, 4),
 “confidence”: round(confidence, 4),
 “entropy”: round(entropy, 4), #
 “flags”: {
 “wrap_left”: wrap,
 “top_present”: top_present,
 “bottom_present”: bottom_present
 },
 }
 params_path = PROCESSING_DATA_DIR / “latest_params.
json”
 json.dump(params, open(params_path, “w”), indent=2)
```

Finally, all relevant parameters are packed into a json and shipped off to the Processing sketch’s data folder for implementation in the sketch.

## 9 Interpretations

The system underpinning the artwork employs five distinct sentence-embedding models: all-MiniLM-L6-v2, all-mpnet-base-v2, paraphrase-multilingual-MiniLM-L12-v2, distiluse-base-multilingual-cased-v2, and sentence-t5-base. (The model is chosen randomly at each input. A version where a simple UI allows the user to select has also been prepared). Although each of these models is designed to map linguistic input into a vector representation suitable for semantic comparison, they differ significantly in architecture, training data, and optimization objectives. As a result, the same word or phrase produces markedly different internal representations across models. When these representations are subsequently constrained and translated into combinations of Chinese radicals, the divergences become visible as formally and conceptually distinct characters. The below sequence of outputs is a simple example of such divergences:

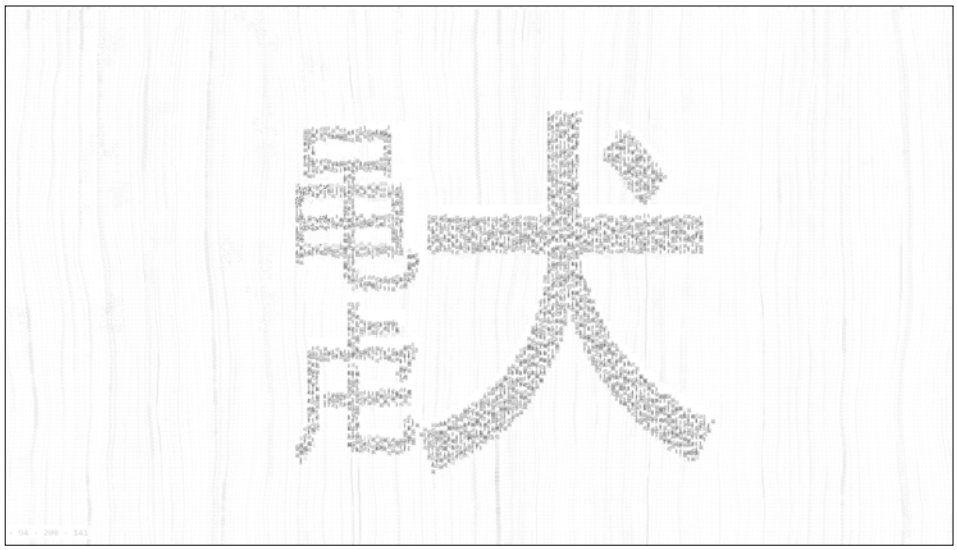

**Fig. 15.** Cat: A turtle-tiger dog (all-MiniLM-L6-v2).

**Fig. 16.** Cat: A mouse-tiger dog (all-mpnet-base-v2)

**Fig. 17.** Cat: A furry mouse tiger (multilingual-MiniLM-L12-v2)

**Fig. 18.** Cat: A dog-mouse tiger (distiluse-base-multilingual-cased-v)

**Fig. 19.** Cat: A dog-bird mouse (sentence-t5-base)

Given the input "*cat*", the five models demonstrate markedly different tendencies in how the concept is internally structured and subsequently articulated through the constrained symbolic system. Across the outputs, the resulting characters variously emphasize attributes such as domestication (*dog*), predatory or feline intensity (*tiger*), smallness (*mouse*), softness or fur (*fur*), avian lightness (*bird*), or even slow, protective embodiment (*turtle*). None of these readings can be said to be incorrect; rather, each reflects a distinct prioritisation of features within the same semantic field. What becomes visible here is not disagreement over meaning, but divergence in *salience* – each model compresses the concept of "cat" according to different internal biases shaped by architecture, training, and representational strategy.

These associations are not culturally neutral. Because such embedding models are trained on large textual corpora, the proximities they learn may reflect not only abstract semantic relations, but also recurring cultural patterns, clichés, and residues of internet language. In this sense, the outputs expose not simply "meaning" in the abstract, but historically and culturally conditioned distributions of salience.

This multiplicity foregrounds a central concern of the work: that meaning is not retrieved intact from language, but assembled through

a process of selective emphasis under constraint. The resulting characters reveal how the system organises the idea of a cat when forced to translate latent representation into symbolic form, thus inviting the viewer to interpret each output not as a depiction, but as a projection – an external trace of internal structure made legible through the shared vocabulary of radicals. In this sense, the characters function less as symbols of an animal and more as snapshots of how different models "think with" the same input when pressed through an identical symbolic bottleneck.

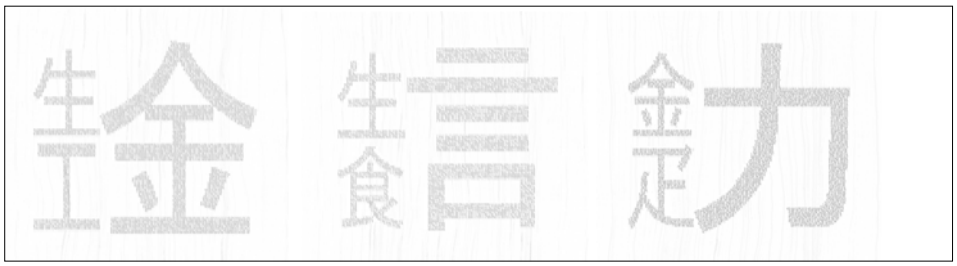

**Fig. 20.** Some further examples based on the author's interpretation: "Money" is described practically as "work life gold" more metaphorically as "life food speech" and as the dramatic "golden cloth power".

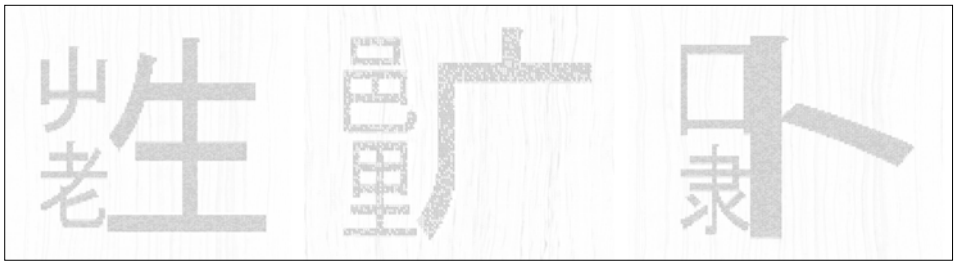

**Fig. 20.** Is "Gentrification" "sprouting old life" (left), an "inner city shelter"(mid) or a "divine slave enclosure"(right)?

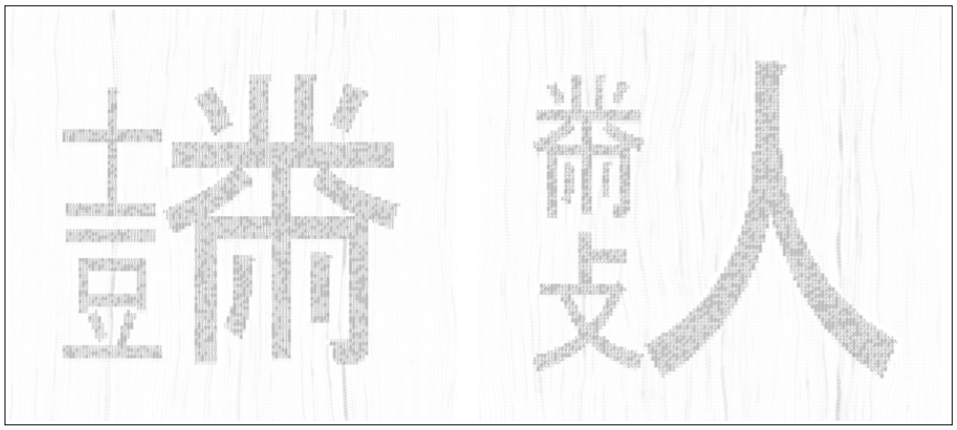

**Fig. 20.** MPNet-base distinguishes between "artisan" and "artist" here. The craft itself ("embroidery") is the focus for the former, while "person" is the primary radical of the latter.

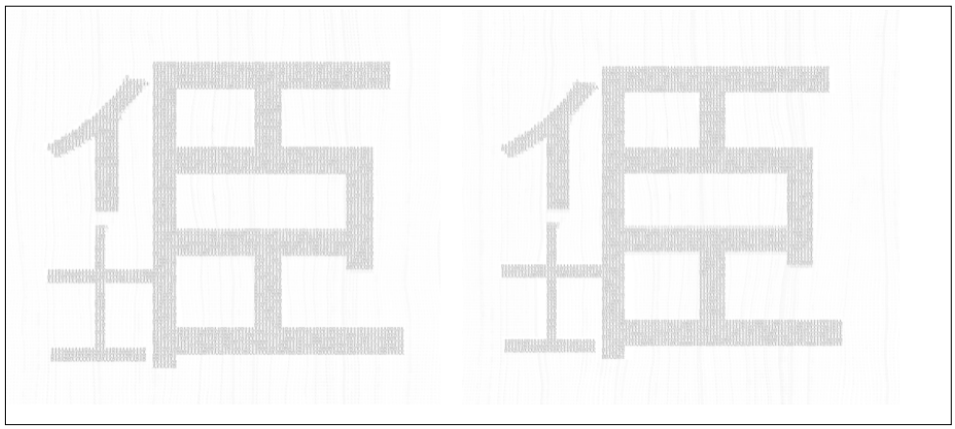

**Fig 23.** Unlike the dramatic differences expressed by the other models, the sentence-t5 transformer sees no reason to differentiate between "politician" (person, scholar, minister) and "profiteer" (same).

## 10 Conclusion

The work presented here set out to experiment with ways in which traces of latent generative structure might be rendered perceptible, if such a domain can meaningfully be said to exist. Through constraint, projection, and symbolic mediation, the project attempts to stage encounters with non-transparent generative structures, allowing them to surface only indirectly, as patterned divergences and provisional forms.

Throughout the paper, interpretation has been foregrounded as a condition of engagement. The variability observed across models, inputs, and symbolic configurations suggests that what is encountered is a shifting field whose contours become visible only under particular pressures. In this respect, the project aligns with depth-psychological approaches in which understanding emerges through attentive observation of patterns over time, instead of direct revelation. The combined radical characters function as sites where such patterns briefly stabilise, inviting reflection without claiming authority. The further examples given above demonstrate this. They are provisional interpretations made solely by the author in response to outputs that remain semantically open. Another viewer would almost certainly read these forms differently, drawing out other associations, emphases, or resonances.

At the same time, the experiment inevitably reflects back upon its human participants. The desire to locate an unconscious within artificial systems mirrors longstanding efforts to understand the forces that shape human creativity, thought, and interpretation beyond conscious control. Contemporary debates surrounding artificial intelligence, which oscillate between utopian expectation and catastrophic anxiety, often overlook this reflexive dimension. The present work suggests that AI may be most instructive not when treated as an autonomous subject or threat, but when approached as a system that exposes the assumptions, symbolic habits, and interpretive frameworks through which we attempt to make sense of intelligence itself.

As such, this project remains deliberately open-ended, proposing one possible method among many for engaging with the opacity of machine-learning systems through artistic practice. Further experimentation – across different symbolic vocabularies, model architectures, material forms, and modes of interpretation – may well reveal other facets of this terrain. If the notion of an AI unconscious retains any value here, it lies in its capacity to orient attention toward what remains unresolved, emergent, and still in the process of becoming intelligible.

## References

**Bender, Emily.**
2024. *Resisting Dehumanization in the Age of 'AI'*. Current Directions in Psychological Science. Available at: https://doi.org/10.1177/09637214231217286

**Guattari, Felix.**
1979. *The Machinic Unconscious: Essays in Schizoanalysis*. Semiotext.

**Hawley, Willis.**
1943. *Oriental Culture Chart #1 The Chinese Radicals*. Hawley.

**Jung, Carl G.**
1964. *Man and His Symbols*. J. G. Ferguson Publishing.

**Lakoff, George, and Mark Johnson.**
1980. *Metaphors We Live By*. University of Chicago Press.

**Legge, James.**
1960. *The I Ching*. Dover Publications
Lindsey, Jack, Wes Gurnee, Emmanuel

**Lindsey, Jack, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, and Craig Citro.**
2025. *On the Biology of a Large Language Model*. Anthropic. Available at: https://transformer-circuits.pub/2025/attribution-graphs/biology.html

**Linge, Ollie.**
2012. *Kickstart Your Character Learning with the 100 Most Common Radicals*. Hacking Chinese. https://www.hackingchinese.com/kickstart-your-character-learning-with-the-100-most-common-radicals/

**LTL Mandarin School.**
2009. *All 214 Chinese Radicals*. LTL Mandarin School. https://ltl-school.info/wp-content/uploads/chinese-radicals.pdf

**McNaughton, William.**
2005. *Reading and Writing Chinese: A comprehensive Guide to the Chinese Writing System*. Tuttle Publishing.

**Monroy, Marco.**
2022. *A complete list of all 214 Chinese radicals and their meaning*. Berlitz. https://www.berlitz.com/blog/chinese-radicals-list

**Simon, Walter.**
1944. *Chinese Radicals and Phonetics with an Analysis of the 1200 Chinese Basic Characters*. Lund Humphries & Co. Ltd.

**Wilhelm, Richard.**
1951. *I Ching or Book of Changes*. Translated by Cary F. Baynes. Routledge & Kegan Paul Plc.