

Article

Not peer-reviewed version

LLMs Judging Architecture: Generative AI Mirrors Public Polls

Nicholas Boys Smith and [Nikos A. Salingaros](#)*

Posted Date: 6 August 2025

doi: 10.20944/preprints202508.0418.v1

Keywords: architectural evaluation; Christopher Alexander; design feedback; emotional design criteria; generative AI; large language models (LLMs); neuroarchitecture



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

LLMs Judging Architecture: Generative AI Mirrors Public Polls

Nicholas Boys Smith ¹ and Nikos A. Salingaros ^{2,3,*}

¹ Create Streets, 81 Lambeth Walk, London SE11 6DX, UK

² Department of Mathematics, The University of Texas, San Antonio, TX 78249, USA

³ Thrust of Urban Governance and Design, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China

* Correspondence: salingar@gmail.com

Abstract

This paper uses large language models (LLMs) to judge among three pairs of architectural design proposals which have been independently surveyed by opinion polls: department store buildings, sports stadia, and viaducts. The tool instructs the LLM to use specific emotional and geometrical criteria for separate evaluations of the image pairs. Those independent evaluations agree almost totally with each other. In all cases, the LLM consistently selects the more human-centric design, and the results align closely with independently conducted public opinion poll surveys. The convergence of AI judgments, neuroscientific criteria, and public sentiment highlights the diagnostic power of AI-based criteria-driven evaluations. This technology provides objective, reproducible architectural assessments capable of supporting design approval and policy decisions. A practical tool that unifies scientific evidence with public preference can promote a more human-centered built environment.

Keywords: architectural evaluation; Christopher Alexander; design feedback; emotional design criteria; generative AI; large language models (LLMs); neuroarchitecture

1. Introduction

1.1. Using AI For Architectural Evaluations

This paper applies generative AI to evaluate designs of buildings and other artificial structures in the environment using scientific criteria. Advanced computational methods have transformed numerous engineering and scientific disciplines. Architects now apply generative AI to create impressive abstract forms [1]. However, the evaluation of the built environment remains rooted in aesthetic judgments carried out by “experts”, which are arguably subjective and have the potential to disagree with both public sentiment and the evidence underpinning human-centric design. The processes of creation and assessment are still detached from each other, so there is no corrective feedback loop.

Convolutional neural networks trained on crowdsourced pairwise comparisons can predict human judgments of urban attributes such as safety, liveliness, and beauty with high accuracy [2]. Deep learning models applied to a large number of online ratings have quantified the scenic beauty of outdoor places, revealing a direct link between environmental features and perceived attractiveness [3]. However, these vision-based approaches often operate as “black boxes”, not integrated with geometrical or neuroscientific theories of human-environment interaction.

Recent interdisciplinary research in neuroarchitecture demonstrates that forms and surfaces exert measurable effects on human cognition, emotion, and well-being, which offers a quantitative basis for design evaluation [4]. Tools based on neuro-architectural assessment can operationalize these scientific insights in practical settings [5]. It would be very convenient to have a technological

tool for acquiring consistently reproducible, criteria-driven evaluations of buildings and urban structures in real time.

This investigation follows pioneering efforts to use generative AI as a tool in evaluating the design of buildings and urban spaces. Initially, large language models (LLMs) have been applied in judging text (not images) that is itself produced by an LLM [6]. This can be done in two distinct ways: (a) direct scoring, where an output is evaluated using specific criteria; and (b) pairwise comparisons in which two outputs are evaluated on relative attributes.

In separate work on visuals, an LLM is asked to evaluate images for their aesthetic characteristics, with surprising success [7]. This direction of research encompasses recent applications of generative AI to judge the degree of objective architectural “beauty”. Pairwise comparison proves to be most useful in relative judgments of objective criteria [8,9]. Recent neuroscientific findings confirm architectural beauty as an objective, biologically-driven phenomenon rather than merely cultural or stylistic preference. Special architectural configurations provoke measurable emotional and physiological responses, forming the basis for reliable, science-based evaluation.

1.2. LLM-Implemented Neuroscientific and Geometrical Criteria.

To address the existing epistemological gaps, this paper introduces a novel generative AI-based evaluation technology that uses large language models (LLMs) for judging architecture. By applying tailored LLM prompts, the tool implements two complementary, evidence and theory-driven sets of criteria for diagnosing and comparing designs, with quantitative results:

1. LLM-Implemented Neuroscientific Criteria. Drawing on peer-reviewed findings in environmental psychology, public health, and neuroaesthetics, we define a “beauty-emotion cluster” comprising ten emotional descriptors: {beauty, calmness, coherence, comfort, empathy, intimacy, reassurance, relaxation, visual pleasure, well-being}. An LLM is prompted to perform pairwise comparisons of images based on these descriptors, producing a normalized preference score for the relative conjectured emotional impact of a design. The set of ten descriptors forming the “beauty-emotion cluster” was introduced by the second author (N.A.S.) [10] in a study of people’s unconscious responses to window shapes and composition.

2. LLM-Implemented Geometrical Criteria. Grounded in Christopher Alexander’s fifteen fundamental properties of living geometry—such as levels of scale, strong centers, nested symmetries, and positive space—this module uses the same pairwise comparison framework to evaluate the presence and dominance of geometrical features that define coherent complexity and informational organization [11]. The prompts include a verbal description of the fifteen fundamental properties (attached at the end of this paper). This comparison generates a preference score that complements the one obtained from the “beauty-emotion cluster”.

Each diagnostic module produces independent, normalized preference ratings. In pilot experiments these align perfectly, demonstrating the internal consistency of emotion- and geometry-based evaluations. Related AI work is listed below (Section 1.3); there is agreement between eye-tracking evaluations and public polls (Section 1.4); the science behind the ten emotional descriptors and the fifteen fundamental properties of living geometry is summarized (Section 1.5).

Alexander’s “Quality Without A Name” — QWAN (better known in computer science than in architecture) characterizes the most intensely humane environments. It is possible to justify this concept with the two sets of criteria used by the LLM diagnostic tool (Section 1.6). It is argued that approval decisions should rely upon criteria-driven diagnostics (Section 1.7).

The paper details the design and implementation of the LLM prompts, reports on the quantitative outcomes of several AI experiments, and describes the methodology and results of three public polls. Three UK projects are analyzed using the AI-based diagnostic tool: Orchard House, the building in Oxford Street, London where the department store Marks and Spencer’s is located [12] (Section 2); a proposed new rugby stadium for Bath [13] (Section 3); and the High Speed 2 trainway viaduct proposal in Solihull outside Birmingham [14] (Section 3). The LLM overwhelmingly chose the more human-centric design over the industrial-modernist alternative in all three cases (Section 4).

This result coincides with independent public polls performed on representative samples of the British public for the three projects conducted for Create Streets by Deltapoll (directed by its chairman, the first author N.B.S.).

To rule out the public debate having any influence on the AI evaluations, the LLM was asked to justify its choice for each of the projects (Section 5). The LLM's self-justification is reproduced in full in Appendix B. Remarkably, the AI analysis is even more categorical in its choices, giving 90% – 100% compared to 70% – 80% in the public polls (Section 6).

This evaluation system represents a technological advance in three key respects. First, the tool leverages LLMs' multimodal reasoning to translate neuroscientific findings and living geometry principles in analyzing images without specialized training or annotated datasets. The technology is ready to use as is, which is a major practical benefit. Second, structuring the evaluation as pairwise comparisons over interpretable semantic criteria overcomes the "black box" limitations of vision-based AI models. The tool relies upon commonly understood emotional/geometric qualities. Third, the platform's modular design allows extension and recursive optimization, enabling it to adapt to emergent insights. Being an open-source tool, any user can fine-tune it. Using generative AI as a theory-grounded diagnostic tool assesses how users perceive the built environment.

The observed alignment across these three evaluation modalities is itself remarkable. Each tool—neuroscientific emotional analysis, assessment of geometrical properties, and public sentiment polling—originates from distinct scientific foundations. Emotional evaluation relies on LLM-driven interpretation of neuro-aesthetic findings and biometric correlation; geometrical evaluation is rooted in informational and mathematical properties; and a public poll captures human preferences via statistical sampling. Their three-way convergence provides mutual reinforcement. The LLMs' ability to merge disparate criteria confirms that intelligent, objective diagnostics can mirror collective human judgments. The agreement strengthens confidence in deploying these AI-based tools for architectural diagnostics.

Significantly, the evaluative model described here does not assess buildings in terms of architectural "style". This AI tool uses prompts that contain no references to classical, historical, industrial, modernist, traditional, or vernacular styles; nor are any images classified or described in stylistic terms. The descriptions are free of value-laden language favoring specific styles. Instead, the LLM evaluates images strictly from primary emotional and geometrical criteria. Misunderstandings on this point undermine the scientific applications of AI-based methodologies to design.

1.3. Some Related Work.

Close to the general thinking of this paper, two recent efforts stand out. Danny Raede has created a website that analyzes an image to find the three out of Alexander's 15 properties that appear most intense [15]. Paralleling the approach adopted here, Raede uses generative AI together with a detailed description of the 15 properties. In a separate development, Bin Jiang has created an online "Beautimeter", also using the 15 properties together with ChatGPT to judge objective beauty [16]. Other work by Jiang tries to apply quantitative methods to architectural judgments [17].

Used as a diagnostic tool, generative AI already knows enough about Alexander's 15 fundamental properties from open sources to evaluate an image for their presence. Such AI experiments differ from the method in this paper that relies on text-based instructions, and which also uses an uploaded description to ensure accuracy. (A detailed description of Alexander's 15 fundamental properties is included below in the Supplementary Materials).

An entirely different approach is trying to identify factors responsible for "place attachment" [18]. That phenomenon occurs from the interaction between people and a particular place through actions, emotions, and thoughts. There is a strong link with the present model, since the AI-based diagnostic test relies upon experienced emotions and separately on the geometry that triggers attachment through visual stimulation.

1.4. Agreement With Public Polls Counters Opinions and Stylistic Biases.

The marketing sector studies the evaluation of product attractiveness, while that important topic so far lies outside the scope of present-day architectural discourse. Going beyond specific questions of polling, it is fascinating to compare how far machine learning (a form of AI) is capable of approximating personal responses obtained via questionnaires [19]. By pushing the architectural debate outside the usual self-referential subjective confines, such tools greatly facilitate the drive towards a more attractive built environment [20,21].

To validate the AI-derived assessments of this paper against human judgments, they are compared to public opinion polls. For example, a February 2024 survey of 1,200 respondents in the UK comparing two alternative façade designs for Orchard House yielded preference scores of (LHS, RHS) = (17 %, 79 %), entirely consistent with the (0 %, 100 %) values returned by both AI-driven diagnostic modules used here. This concordance underscores the mutual support among neuroscientific theory, geometrical theory, and empirical public sentiment.

A recent turn in architectural criticism circumvents the authority of established aesthetic movements by comparing eye-tracking technology and public surveys. Two independent analyses of similar pairs of US Federal Buildings validated results of an independent Harris Poll of public opinion. Eye-tracking experiments and eye-tracking simulation software independently verified the survey's results [22–24]. That survey revealed overwhelming public preference for US federal buildings in traditional versus modernist styles [25,26]. Mapping unconscious visual attention overcomes architectural beliefs and stylistic attitudes.

1.5. Scientific Basis for the Emotion/Geometry Criteria.

The present model's ten emotional descriptors were chosen following a systematic review of the neuro-architectural literature. Psycho-physiological studies employing EEG, galvanic skin response, and salivary cortisol measurements link emotional responses such as calmness, comfort, and visual pleasure to spatial stimuli [27,28]. These descriptors reflect high positive-valence states linked to beneficial health outcomes in hospital and restorative environments. The proposed "beauty-emotion" cluster aligns with biophilic design principles identified in empirical studies [29–33].

A reader should therefore not mistakenly assume that the emotional descriptors were justified by a single study of window shapes in which they were introduced: they are instead supported by comprehensive psycho-physiological evaluation research. The ten descriptors may appear conceptually redundant and measure overlapping variance — yet their apparent construct overlap intentionally captures emotional nuances for evaluation. Eventually, correlation matrices between the 25 combined criteria will be calculated to identify redundant measures, but that is not the aim of the present paper.

Similarly, Alexander's fifteen fundamental properties were selected due to their empirical validation in both built and natural contexts. Cross-disciplinary analyses reveal that structures exhibiting these properties correlate geometric coherence with human psychological well-being [34–36]. Empirical peer-reviewed studies demonstrate their measurable well-being outcomes — yet more controlled experiments are needed measuring physiological responses to designs varying in Alexander's properties. Although other geometric frameworks could be employed in an LLM prompt, they would lack the synergistic, multi-scale organization shown to support cognitive and emotional health.

More important, violating living geometry leads to negative physiological reactions [37,38]. While the 15 fundamental properties inherently describe features more prevalent in classical and traditional architectures than in modernist designs, this result should not be misinterpreted as a systematic bias. It is a consequence of deliberate stylistic choices made by those who introduced early modernism at the beginning of the 20th century, whereas the 15 properties are justified from their mathematical content and psychological effects on people.

The fifteen properties offer a convenient, operational approach to judging living geometry. There is a lot of mathematical information that lies embedded beneath these simple descriptors, and

those complex structures have been analyzed in depth by the second author (N.A.S.) [39–41] (who was the principal editor of Alexander’s 4-volume book *The Nature of Order*). This theoretical grounding ensures that the chosen criteria used in this AI tool are neither arbitrary nor stylistic but are rooted in robust geometrical and neuroscientific evidence.

The ten emotional descriptors and the fifteen fundamental properties of living geometry are conceptually distinct (though neurologically linked) methods of analysis. The fifteen fundamental properties measure a form’s intrinsic, visible structural features. Those are assessed from the geometry of the design itself, independent of an observer. By contrast, the ten emotional descriptors are not intrinsic to the structure, but rather represent the affective and physiological responses of a person interacting with that structure. These 25 combined criteria do not represent statistically independent measures, for the following reason.

There is a causal relationship between the two sets of criteria: the 15 geometric properties (along with other sensory input) trigger the 10 emotional responses through our hardwired neurobiological systems. Brains evolved in natural environments to interpret the living geometry that Alexander identified. Built environments trigger our nervous systems to respond with an unconscious survival assessment. The emotional descriptors function as biofeedback proxies: mediated effects through our neurophysiology. It is diagnostically valuable to test with both sets of criteria independently, even though they yield similar results.

1.6. Alexander’s QWAN – *The Quality Without A Name*.

In 1979, Alexander defined the “Quality Without A Name” (QWAN) in *The Timeless Way of Building* [42] as the ineffable attribute of the most emotionally-resonant and humane places. This special quality is weak or missing entirely from impersonal and sterile environments. Alexander described the QWAN as these seven qualities combined: {alive, whole, comfortable, free, exact, egoless, eternal}. He did not develop the QWAN further, and went on to discover the fifteen fundamental properties. Although architects largely ignored the concept of the QWAN, computer scientists and software engineers recognized that it describes well-designed systems.

The combined use of emotional and geometrical criteria in the present model offers an implementation of the QWAN. The 15 fundamental properties of living geometry are inherent structural features that generate the conditions for the QWAN, while the 10 emotional descriptors represent the embodied response that arises when a person perceives such conditions. The 7 QWAN attributes are phenomenological and recognized only through deep feeling and intuition. However, the LLM diagnostic framework captures the QWAN’s essential attributes – hence succeeds in measuring it. The present AI model therefore judges the conditions under which the QWAN emerges.

1.7. *Criteria-driven Diagnostics Should Influence Approval Decisions*.

Criteria-driven diagnostics pave the way for decision-support systems in heritage conservation, regulatory review, and urban planning, where built structures can be compared and scored according to scientifically-grounded criteria. Importantly, this technology decouples the assessment process from aesthetic fashions, historical or personal biases, and even the limitations of manual surveys. Those who work comfortably within the present system might find the proposed change controversial, however.

Introducing AI-based diagnostics into design practice poses a cultural and institutional challenge, as many practitioners view both evidence-based methods and public opinion as subordinate to aesthetic authority. To address such concerns, the technology must be understood as a decision-support system that augments rather than replaces professional expertise. Quantified emotional and geometrical feedback enhances design narratives, supports client engagement, and aligns with emerging public health data. Scientific validation couples with improved occupant well-being and popular market appeal, making AI-based evaluation tools an indispensable resource for evidence-based, human-centered architecture.

A downloadable description of Alexander's 15 fundamental properties of living geometry is included in the Supplementary Materials at the end of this paper.

Appendix A discusses the consistency of the two LLM evaluations — emotional and geometrical criteria — and performs reliability assessments both for a single LLM, and across different LLMs.

Appendix B includes the self-justification that ChatGPT gave affirming that it did not draw from open-source data on the public debates around the three examples evaluated here.

2. Two AI Experiments Use Large Language Models to Judge Architecture.

2.1. AI Experiment 1. Comparative LLM evaluation from the “Beauty-Emotion Cluster”

This generative AI experiment uses the large language model (LLM) ChatGPT to perform two independent comparative evaluations of the two buildings shown in Figure 1. First, the LLM estimates the combined emotional feedback of a person who experiences either building in the actual physical setting. This gives a preference for one building over the other based on the “beauty-emotion cluster”. The initial prompt for the AI experiment was entered as a new Chat, to avoid interference with and memory from any previous activity in the LLM.

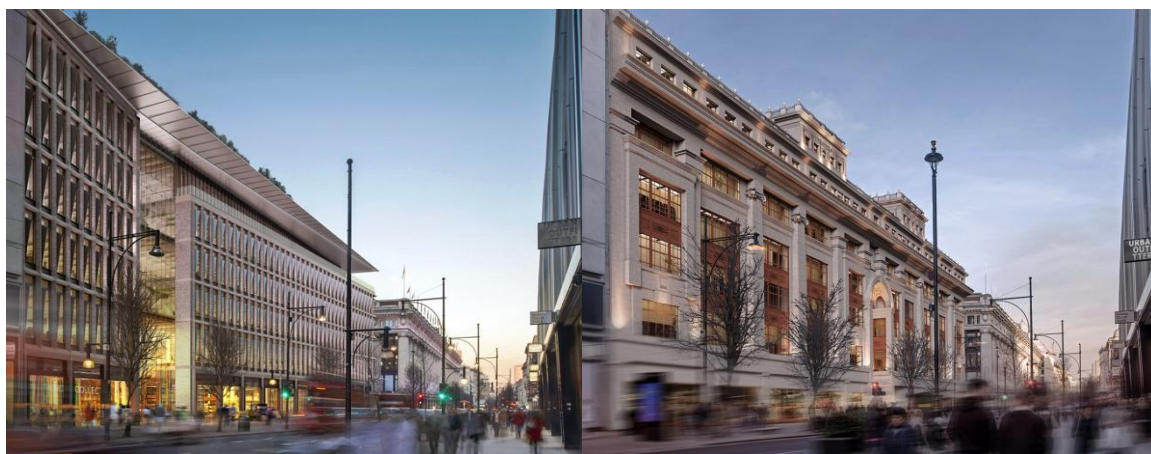


Figure 1. Two images of buildings of the same size in the same urban setting. Images by Pilbrow & Partners (LHS) and Francis Terry and Associates (RHS), used with permission.

Second, the LLM performs a comparative evaluation of specific geometrical properties that define what is known as “living geometry” and described by Alexander's fifteen fundamental properties. “Living geometry” represents the common properties of stable complex systems, whereas the “beauty-emotion cluster” is a set of responses that human evolution developed to recognize living geometry. The two separate experiments are normalized to a percentage value, and in this example, turn out to agree completely.

Prompt to ChatGPT 4.5: “Use the set of ten qualities (“beauty–emotion cluster”) that elicit a positive-valence feeling from a person while physically experiencing a built structure, to investigate the two uploaded pictures of similar buildings:

- beauty;
- calmness;
- coherence;
- comfort;
- empathy;
- intimacy;
- reassurance;
- relaxation;
- visual pleasure;

well-being.

Evaluate the conjectured relative emotional feedback by comparing the two images in a binary preference (1 for the preferred image and 0 for the rejected image for each of the 10 qualities) to give a preference for one over the other. The sum of the values for each image should be 10, so multiply by 10 so the sum is out of 100%. Give the answer as (LHS, RHS) where $LHS + RHS = 100$."

Chat GPT: "Evaluating the two buildings based on the listed emotional qualities, the relative binary preference scores (1 for preferred, 0 for rejected) are as follows:

Table 1. Relative scores for the emotional qualities in Figure 1. Table by ChatGPT.

<i>Emotional Quality</i>	<i>Left-hand Image (LHS)</i>	<i>Right-hand Image (RHS)</i>
Beauty	0	1
Calmness	0	1
Coherence	0	1
Comfort	0	1
Empathy	0	1
Intimacy	0	1
Reassurance	0	1
Relaxation	0	1
visual pleasure	0	1
well-being	0	1

Total Relative Preference: (LHS, RHS) = (0%, 100%). Thus, the overall preference clearly favors the building in the second image (RHS)."

2.2. AI Experiment 2. Comparative LLM evaluation from the qualities of "living geometry".

For this second experiment, it was necessary to upload the document "Detailed description of Christopher Alexander's 15 fundamental properties" (available in the Supplementary Materials) along with the prompt given below. Christopher Alexander presented his geometrical properties in Book 1 of *The Nature of Order* [43]. This condensed description comes from the Appendix of an article on living geometry by the second author (N.A.S.) [44].

Prompt to ChatGPT 4.5: "Do a similar evaluation of the same two images of buildings, this time using the 15 criteria uploaded as Alexander's Fifteen Fundamental Properties of living geometry. The relative comparison should come out again as a set of numbers (LHS, RHS), where LHS = total score for the relative presence (dominance) of the properties in the LHS image, and RHS = total for score for the relative presence (dominance) of the properties in the RHS image. Score the pair of images as follows: if one property is clearly dominant in one of them, give a 1 to it and 0 to the other. If both images have comparable degrees of one property, or the difference is very small, give a 0 to both. For this reason, the totals could come out to be $LHS + RHS < 15$. Whatever this sum is, normalize it so that the two numbers add up to 100%."

ChatGPT: "Here is the evaluation comparing the two images using Alexander's 15 Fundamental Properties, clearly indicating dominance (1) or absence/ambiguity (0) for each criterion:

Table 2. Relative scores for the geometrical qualities in Figure 1. Table by ChatGPT.

<i>Fundamental Property</i>	<i>Left-hand Image (LHS)</i>	<i>Right-hand Image (RHS)</i>
1. Levels of Scale	0	1
2. Strong Centers	0	1
3. Thick Boundaries	0	1
4. Alternating Repetition	0	1
5. Positive Space	0	1
6. Good Shape	0	1
7. Local Symmetries	0	1

8. Deep Interlock/Ambiguity	0	1
9. Contrast	0	1
10. Gradients	0	1
11. Roughness	0	1
12. Echoes	0	1
13. The Void	0	1
14. Simplicity/Inner Calm	0	1
15. Not-Separateness	0	1

The evaluation clearly demonstrates complete dominance by the building shown in the right-hand image (RHS), which strongly exhibits Alexander's properties of living geometry. The left-hand image (LHS) lacks distinct evidence of these essential properties. Final Normalized Preference: (LHS, RHS) = (0%, 100%)".

3. Two More Examples: A Stadium and a Railway Viaduct.

3.1. Analysis of designs for a proposed Rugby stadium in Bath, UK.

The same prompts to ChatGPT given in the previous section were used for evaluating the dominance of emotional qualities and properties of living geometry for the following pairs of images. Only the answers are included here. The format and method of answering differs each time and contains interesting details of the LLM's reasoning.

A comparative analysis of two alternative proposed designs for a rugby stadium is performed on the images in Figure 2. The identity and location of the project were not communicated to the LLM, to keep the analysis as unbiased as possible. The point of this exercise is to show that human-centered design is independent of scale. Applying this diagnostic tool to large structures — not buildings — justifies its universal application. It also questions the common assumption that such projects must automatically adopt an industrial-modernist vocabulary.

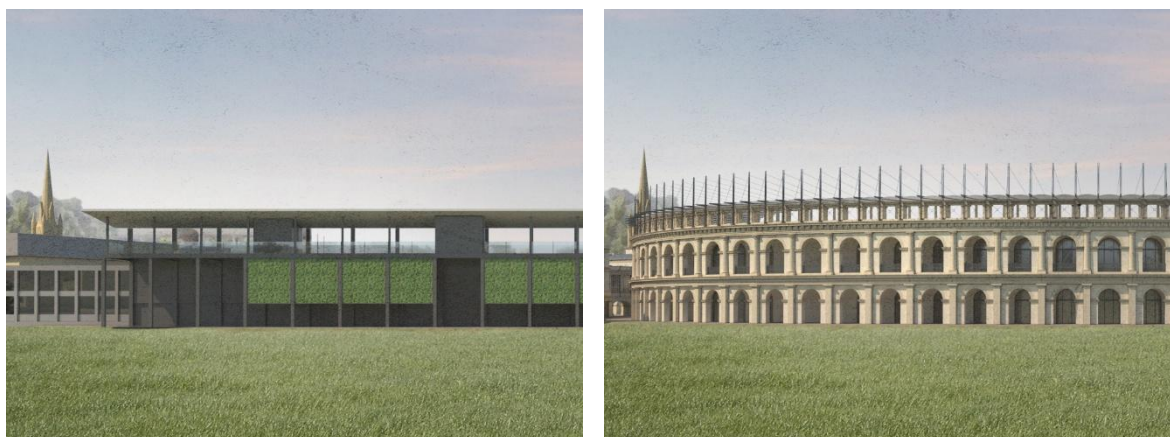


Figure 2. Two stadia of comparable size in the same rural setting. Images by Bath Rugby Club (LHS) and Apollodorus (RHS), used with permission.

Chat GPT-4o: "Using the "beauty-emotion cluster" to compare the two images of the rugby stadium design proposals, I assessed the ten emotional qualities based on the likely instinctive response of a person physically experiencing each structure. Here's the binary scoring:

1. Beauty: LHS: 0 (plain, utilitarian appearance), RHS: 1 (ornamented, harmonious form).
2. Calmness: LHS: 1 (minimalist, quiet), RHS: 0 (visually busy, more stimulating).
3. Coherence: LHS: 0 (flat and inconsistent rhythm), RHS: 1 (structured colonnade, classical order).

4. Comfort: LHS: 0 (hard edges, few cues of warmth), RHS: 1 (human-scale arches, traditional cues).

5. Empathy: LHS: 0 (abstract, machine-like), RHS: 1 (reads as built with care for human experience).

6. Intimacy: LHS: 0 (large, industrial proportions), RHS: 1 (arcade and window rhythm invite engagement).

7. Reassurance: LHS: 0 (anonymous, could be anything), RHS: 1 (rooted in familiar typologies).

8. Relaxation: LHS: 0 (institutional aesthetic), RHS: 1 (inviting, layered detail).

9. Visual Pleasure: LHS: 0 (flat planes, uniformity), RHS: 1 (rich texture, shadow play, ornament).

10. Well-being: LHS: 0 (uninspiring, lacks identity), RHS: 1 (offers delight and sense of place).

Total Scores: (LHS, RHS) = (1, 9), converted to percentages: (LHS, RHS) = (10%, 90%)."

The second step is to analyze this pair of images using their correspondence to living geometry, as defined by Alexander's Fifteen Fundamental Properties.

ChatGPT o4-mini-high: "Below is the comparison of the two stadium façades against Alexander's Fifteen Fundamental Properties. A score of 1 indicates clear dominance of that property in the image; 0 indicates no dominance or near-equivalence. The raw totals are then normalized to sum to 100%.

Table 3. Comparison of the two stadium façades using Alexander's Fifteen Fundamental Properties. Table produced by ChatGPT.

#	<i>Fundamental Property</i>	<i>LHS</i>	<i>RHS</i>
1	Levels of Scale	0	1
2	Strong Centers	0	1
3	Thick Boundaries	0	1
4	Alternating Repetition	0	1
5	Positive Space	0	1
6	Good Shape	0	1
7	Local Symmetries	0	1
8	Deep Interlock and Ambiguity	0	1
9	Contrast	0	1
10	Gradients	0	0
11	Roughness	0	1
12	Echoes	0	1
13	The Void	0	1
14	Simplicity and Inner Calm	0	1
15	Not-Separateness	0	1
	Raw Total	0	13
	Normalized Total	0 %	100 %

Final Normalized Preference: (0 %, 100 %)."

3.2. A proposed viaduct for the High Speed 2 (HS2) trainway in Solihull, UK.

A comparative analysis of two proposed designs for viaducts is performed on the images shown in Figure 3. Again, any information that might identify the project hence lead the LLM to be influenced by the public debate or poll results was excluded from the prompt.



Figure 3. Two images of viaducts of the same size in the same rural setting. Images by HS2 (LHS) and Create Streets Ltd. (RHS), used with permission.

Chat GPT-4o: “Using the “beauty-emotion cluster” of 10 experiential qualities, here is an evaluation of the two viaduct designs you provided: Viaduct One (LHS): Modern concrete viaduct with minimalistic piers and industrial detailing. Viaduct Two (RHS): Brick-arched viaduct with classical symmetry and ornamentation.

Table 4. Evaluation Using the 10 Qualities. Table produced by ChatGPT.

<i>Quality</i>	<i>Preferred Image</i>	<i>Reason</i>
Beauty	RHS (1)	The brick arches offer proportion, texture, and ornamentation
Calmness	RHS (1)	The rhythmic arches and warm material palette are visually soothing
Coherence	RHS (1)	The classical structure is more legible and harmonious in design
Comfort	RHS (1)	The design feels more human-scaled and less imposing
Empathy	RHS (1)	The brick evokes historical and emotional familiarity
Intimacy	RHS (1)	Arched forms and materials feel less alienating and more grounded
Reassurance	RHS (1)	The traditional style communicates timelessness and durability
Relaxation	RHS (1)	Less aggressive structure, better suited to natural landscape
Visual Pleasure	RHS (1)	Arches and detailing provide visual interest and delight
Well-being	RHS (1)	The overall character blends more naturally with a public realm

Binary Preference Summary. LHS (Viaduct One): 0/10, RHS (Viaduct Two): 10/10.

Final Score (out of 100%): (LHS, RHS) = (0%, 100%).

This result suggests that, from the perspective of emotional experience and human-centred design, the brick-arched viaduct (RHS) overwhelmingly outperforms the modernist concrete version (LHS) across all ten emotional criteria.”

The next AI experiment evaluated the comparative presence of the qualities of living geometry in Figure 3, as judged from Alexander’s Fifteen Fundamental Properties.

ChatGPT o4-mini-high: “Below is the comparative evaluation of the two viaduct designs against Christopher Alexander’s Fifteen Fundamental Properties. A score of 1 indicates clear dominance of that property in that image; 0 indicates absence of dominance or near equivalence. The raw totals are then normalized so that they sum to 100 %.

Table 5. Comparative evaluation of the two viaduct designs against Christopher Alexander’s Fifteen Fundamental Properties, using ChatGPT. Table generated by ChatGPT.

#	<i>Fundamental Property</i>	<i>LHS</i>	<i>RHS</i>
1	Levels of Scale	0	1
2	Strong Centers	0	1

3	Thick Boundaries	0	1
4	Alternating Repetition	0	0
5	Positive Space	0	0
6	Good Shape	0	1
7	Local Symmetries	0	0
8	Deep Interlock and Ambiguity	0	1
9	Contrast	0	1
10	Gradients	0	0
11	Roughness	0	1
12	Echoes	0	1
13	The Void	0	1
14	Simplicity and Inner Calm	1	0
15	Not-Separateness	0	1
	Raw Total	1	10
	Normalized Total	9 %	91 %

Final Normalized Preference: (9 %, 91 %)."

4. Consistent Results From Evaluating the Three Building Proposals.

Table 6 summarizes the separate AI-based evaluations based on the emotional indicators and properties of living geometry for the three pairs of examples. The public survey results are also included for comparison. Using different ChatGPT versions across experiments does not undermine inter-version reliability, but illustrates independence of the model from any specific LLM (this crucial point is elaborated in Appendix A). It is important to note that the name of the project was not given to the LLM when it performed its analysis. This deliberate omission avoided bias from the LLM picking up the public survey results or opinions from the associated debate (see the discussion in Section 5 and Appendix B).

Table 6. Summary of the three projects shown in Figures 1, 2, and 3.

<i>Project</i>	<i>Emotional Score</i>	<i>Geometrical Score</i>	<i>Public Survey</i>
Orchard House	(0%, 100%)	(0%, 100%)	(17%, 79%)
Bath Rugby Stadium	(10%, 90%)	(0 %, 100 %)	(28%, 72%)
HS2 Viaduct	(0%, 100%)	(9 %, 91 %)	(28%, 69%)

There is complete agreement between the single-trial AI-based analysis and the public poll in choosing the human-centered design on the RHS (Appendix A investigates the data variation expected in repeated trials). What is remarkable is that in all three cases, the AI result is even more categorical than the opinion survey. Possible reasons for this phenomenon will be discussed in the following sections. The LLM is telling us something important about the human impact of these designs that is not yet fully appreciated.

The binary scoring methodology was chosen as the simplest initial implementation of the idea, which does not compromise measurement validity. Certainly, nuanced gradations — such as a Likert scale — essential for a more complex assessment can be easily incorporated into future models. Convenience and the need for a simple scoring format (to communicate to authorities) forces the raw binary scores into a percentage, lowering the accuracy of detecting when both designs perform equally poorly or well on the evaluation criteria.

This innovative approach to design evaluation using large language models with the “beauty-emotion cluster” and Alexander’s fifteen fundamental properties as measurement instruments still needs empirical validation. The sample size of three architectural pairs and the absence of appropriate control conditions provides insufficient statistical weight for conclusions. Nevertheless, these preliminary findings are meant to spur developing design validation in this direction. The

methodology does not rely on single-trial evaluations, even though, for simplicity, the paper uses one evaluation per example. The LLM tool's consistency and reliability are discussed in Appendix A.

The purpose of this pilot project was to present the basic notion of AI-based design diagnostics, applied to some real examples. The authors admit to conjoining causal claims with correlational results, yet this point will need to be clarified by future research. A discussion on "method limitations" is placed in Appendix A rather than in the body of the paper, since generative AI is developing so fast that outstanding questions may be answered quite soon.

It would have been ideal to have appropriate control conditions to validate the evaluation methodology against established architectural assessment tools. But the problem is that, at present, there is no objective standard assessment method to compare with [45–49]. This poses a huge dilemma since expert ratings from dominant practice have historical and socio-political biases, and to analyze them is beyond the scope of the present paper.

5. The Three Building Surveys Did Not Bias This AI-Based Analysis.

The three studies undertaken here have a prior history in the UK, where public polls were conducted to judge preferences in each case (see Section 6, below). Before discussing those surveys, it is important to dispel a question that arises about the impartiality of the LLM-based evaluation. Since the LLM has access to open-source data, could it be influenced by the public poll? In that case, whatever result it produces will not be due strictly to scientific criteria as the model promises.

This question was settled by asking the LLM itself to justify its judgment in each of the three cases. ChatGPT gave clear and detailed answers that categorically deny any influence of the public polls on its results. The unedited responses are reproduced here as Appendix B. Some readers of early versions of this paper did not believe ChatGPT's self-assessment as a control measure because LLMs lack metacognitive awareness — which is not true, since the latest models display partial metacognition [50]. The authors are not worried about this point, and those who are still skeptical are welcome to investigate possible hidden bias.

6. Discussion: Why Criteria-Driven Architectural Evaluations May Give a Clearer Result than Even Public Opinion Polls.

The public opinion surveys previously commissioned by Create Streets use a strict Visual Preference Survey methodology. This framework carefully controls for extraneous factors to permit a fair understanding of human preferences. Such visual preference surveys have very consistent results. Around 70-80% of the public consistently prefer more traditional and human-centered designs. In the three visual comparisons reviewed here, 79%, 72% and 69% of the public preferred the more human design. Remarkably, these findings are consistent across all socio-economic demographic segments of social strata, geography, race, sex, income, and political opinion. They are also true in other surveys in the UK and in the US not reviewed here [51,52].

Clearly, public opinion surveys of visual preference comparisons perform an important role that has been lacking for much of the last century. They demonstrate strong but not unanimous, cross-party and widespread support for more humane architecture and places. Public opinion provides important and useful feedback on architecture. It is also immediately comprehensible and usable by politicians and decision-makers.

The public survey of (LHS, RHS) = (17%, 79%) offers a representative picture of people's subjective preferences of the department store in Oxford Street. Both the AI-based emotional and geometric analyses gave a 100% preference to the RHS image. Context can sway humans and since the RHS image is a sensitive upgrade of the historic Orchard House, some people might prefer it out of heritage value or nostalgia, which would boost the RHS score — not the LHS. This does not explain the AI-driven analyses of a 0% score for the LHS image, which was deliberately performed without historical reference.

Abstract analyses using specific emotional/geometric criteria and public polls are not redundant tools, contrary to what people might assume. In fact, an analysis that uses objective criteria such as implemented in this paper is arguably superior to a public poll. People weigh features differently and have cultural/emotional attachments that contradict objective criteria. Unlike a human respondent, however, an LLM is immune to ideology, prestige bias, or social pressure.

Human responses to designs are not purely subjective or socially conditioned — they are deeply rooted in our neurobiology [53]. Our brains evolved in natural environments rich with face-like symmetries, fractal patterns, and organic forms, which provided essential cues for survival [54,55]. The geometry itself has the potential for a therapeutic effect, likely because our visual system evolved to process the special complexity of natural environments efficiently and pleasurably. This informational criterion translates directly into better physical health metrics (lowered physiological stress and a calmer bodily state).

Violating these innate emotional and geometric criteria triggers anxiety and stress. Forms aligned with evolutionary expectations tend to feel comfortable, whereas radically novel or abstract structures can elicit unconscious “warning” responses of danger. Even if someone claims to prefer such structures, that person’s autonomic nervous system will show greater signs of stress, whether they consciously register it or not [56]. Ideological or stylistic conditioning, especially in architectural education and media, trains some people to claim to prefer geometries that trigger physiological stress responses. Researchers have therefore found a consistent “design disconnect” between the preferences of those educated in architecture schools and the wider public in several countries [57–59].

Design evaluation should be using human-centered criteria grounded in biology, because those criteria predict real health outcomes. A simple popularity poll cannot capture this fact. Public opinion can be swayed by familiarity, fashion, or intellectual ideology, and a small subset of people may even insist they like buildings that, unbeknownst to them, cause physiological stress. The mere exposure effect [60] causes people to prefer what they’ve seen often, even if it is harmful or stress-inducing. However, liking something because of social learning does not immunize one’s body against its effects.

Criteria-driven evaluations grounded in neuroscience and physiology could therefore be more accurate than opinion polls for assessing architecture. They capture the deep, unconscious, human responses that determine whether a space nourishes us or wears us down. They identify designs that satisfy people’s neurological needs (regardless of background) and flag designs that consistently provoke latent discomfort. Public polls will always have a biological margin of error, because people can be poor judges of the sources for their own latent stress. That remaining minority, for idiosyncratic reasons, say they favor unhealthy design.

7. Conclusions

A generative AI system, prompted with neuroscientific and geometric design criteria, consistently selects the human-centered alternative. In three real-world building proposals—a department store renovation (Orchard House), a stadium (Bath), and a railway viaduct (High Speed 2 in Solihull outside Birmingham)—the generative AI model consistently rejected the mainstream architectural proposals and chose the more popular and traditional design. Crucially, the AI’s evaluations were impartial: it was not told which designs were “modern” or “traditional”, nor did it know about the political or professional debates surrounding those projects. Yet in all three instances, the LLM gave overwhelming scores (90–100%) in favor of the more biophilic, historically-informed design.

Developing an AI-based diagnostic tool reinforces the robust scientific foundation of objective beauty as a measurable biological phenomenon. The convergence between generative AI evaluations and public sentiment can be most clearly understood through shared, underlying biological reactions to architectural forms and surfaces. Diagnostic criteria grounded in neuroscience and geometry (universal and consistent across diverse populations) offer a significant advantage over purely subjective aesthetic judgments.

The evaluative model of this paper reveals a looming confrontation between two powerful, yet fundamentally divergent, epistemological paradigms shaping the built environment. On one side, industrial modernism has long championed aesthetic paradigms that are largely subjective, resistant to empirical validation, and indifferent or even hostile to both neuroscientific evidence and public preferences. On the other side, the rapidly emerging field of artificial intelligence now offers data-driven, objective assessments of architectural designs, grounded in empirical research and human-centered criteria.

Mainstream architecture, particularly as taught in elite schools and practiced in high-profile firms, operates on deep-seated stylistic assumptions that reject popular preference as uninformed. Generative AI used in the present model does the exact opposite in agreeing with independent public opinion polls. The most powerful contribution of this paper is not just technical—it is epistemological. If LLMs trained on a vast corpus of scientific knowledge consistently reject design proposals favored by the mainstream profession, this exposes a contradiction between those products and the empirical grounding of human well-being. The LLM, unburdened by ideology, reveals what human beings actually need from buildings.

Supplementary Materials: The pdf file “Detailed description of Christopher Alexander’s 15 fundamental properties” The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualization, N.A.S. and N.B.S.; methodology, N.A.S.; software, N.A.S. and N.B.S.; validation, N.A.S. and N.B.S.; formal analysis, N.A.S. and N.B.S.; investigation, N.A.S. and N.B.S.; resources, N.A.S. and N.B.S.; writing—original draft preparation, N.A.S.; writing—review and editing, N.A.S. and N.B.S.; visualization, N.A.S. and N.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All relevant data are included in this paper.

Acknowledgments: The diagnostic tool uses generative AI to obtain results, in particular different versions of ChatGPT. AI-generated text and tables are shown in quotes. This paper was triggered by an interview of the first author (N.B.S), Theory of Architecture #27, 9 July 2025 where Bruce Buckland asked for “An AI or some form of algorithm that measured the visual complexity of a façade system, and a requirement for that value to be over a certain level” (32:00 in the video).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
LLM	Large Language Model
LHS	Left Hand Side
RHS	Right Hand Side
QWAN	Quality Without A Name

Appendix A. LLM Reliability Assessment and Test-Retest Consistency.

Appendix A.1. Emotional criteria.

A reliability assessment helps to identify the consistency and dependability of the LLM-based architectural evaluation tool. The inherently stochastic nature of generative AI systems means that identical prompts can yield different responses. A test-retest reliability analysis is conducted by querying ChatGPT-4o ten times with identical prompts for the same architectural pair in Figure 1. The standard deviation from the mean is the simplest consistency measure. (A test-retest reliability

coefficient r is not useful here because successive trials vary randomly and are not supposed to converge). This version of ChatGPT was chosen for this test because it is the most widely used at the time of writing.

A slightly modified prompt is employed, and each query is entered as a new chat:

“Use the set of ten qualities {beauty, calmness, coherence, comfort, empathy, intimacy, reassurance, relaxation, visual pleasure, well-being} (“beauty–emotion cluster”) that elicit a positive-valence feeling from a person while physically experiencing a built structure, to investigate the two uploaded pictures of similar buildings. Evaluate the conjectured relative emotional feedback by comparing the two images in a binary preference (1 for the preferred image and 0 for the rejected image for each of the 10 qualities) to give a preference for one over the other. The sum of the values for each image should be 10. Give the answer as (LHS, RHS).”

This LLM produced the following results when evaluating for the emotional criteria:

(LHS, RHS) = (0, 10) seven times and (1, 9) three times.

Mean = (0.3, 9.7) and standard deviation = (0.46, 0.46).

The lesson for researchers is that, to improve reliability, an evaluation should be repeated several times.

Extensive trials indicated that the best model to use for this comparative analysis is the more advanced ChatGPT 4.5, not 4o, which is what the body of the paper quotes for the emotional evaluation. The reason for this choice is that the detailed explanations given by ChatGPT 4.5 proved to be incisive and unbiased, and not because the numbers agreed with the authors’ expectations. According to ChatGPT, version 4.5 is slower but more deterministically reliable in structured scoring tasks than 4o, because 4.5 has lower stochastic entropy and is better aligned with fixed evaluation frameworks.

The second reliability assessment checks whether different LLM versions, and distinct LLMs, will produce comparable results. Inter-version reliability is established by comparing evaluations across ChatGPT-4o, o3, o4-mini, o4-mini-high, 4.5, and 4.1 using the image set in Figure 1. The evaluation trial is extended to include the LLMs Copilot and Perplexity (neither of which has its own AI engine but relies on those of other LLMs). The following numbers will of course change over repeated runs, so this is merely an indication of what to look for in a reliability check.

Single-trial results from ChatGPT-4o, o3, o4-mini-high, 4.5, 4.1, Copilot, and Perplexity were all equal for this case: (LHS, RHS) = (0, 10), whereas ChatGPT o4-mini scored (2, 8).

Mean = (0.25, 9.75) and standard deviation = (0.66, 0.66).

Claude Sonnet 4, Gemini 2.5 Pro, and Kimi K1.5 — LLMs with their own AI engines — gave inconsistent results with the above simple prompt. This was due to their conjecturing of effects for the emotional qualities that amounted to speculation. There is an evident training data bias that conflates aesthetic and emotional criteria. Those LLMs’ detailed explanations revealed that their results were not based strictly on documented psychological feedback but were influenced by opinions on contemporary aesthetics and styles. To use those LLMs, a more detailed prompt will be necessary to prevent the LLM from picking subjective opinions instead of searching through scientific data. An experiment with Gemini 2.5 Pro gave better results, as detailed below in Appendix A.3.

This exercise in response consistency is not a rigorous reliability test for the emotional evaluation module. It simply points out what researchers must do in a systematic manner to validate this model for future investigations. Another important point that came out of this is that distinct LLMs answer questions differently, by drawing upon different sources that may indeed be biased. For this reason, it is essential to ask the LLM for a detailed justification for each number in the evaluation and to check this for impartiality.

Appendix A.2. Geometric criteria.

The test-retest reliability analysis was repeated for the 15 fundamental properties by querying ChatGPT-4o ten times with an identical prompt for the same architectural pair in Figure 1. Each query

was entered as a new chat. A slightly modified prompt was used this time, along with the descriptive list of the 15 properties:

“Evaluate these two images of buildings, using the 15 criteria uploaded as Alexander’s Fifteen Fundamental Properties of living geometry. The relative comparison should be presented as a set of numbers (LHS, RHS), where LHS = total score for the relative presence (dominance) of the properties in the LHS image, and RHS = total score for the relative presence (dominance) of the properties in the RHS image. Score the pair of images as follows: if one property is clearly dominant in one of them, give a 1 to it and 0 to the other. If both images have comparable degrees of one property, or the difference is very small, give a 0 to both. For this reason, the totals could come out to be $LHS + RHS < 15$.”

ChatGPT-4o produced the following results when evaluating the geometrical criteria ten consecutive times (listed here in no particular order):

(LHS, RHS) = (0, 15), (0, 13), (1, 13), (2, 12) four times, (3, 10) twice, (3, 11).

Mean = (1.8, 12.0) and standard deviation = (1.1, 1.4).

The second reliability assessment compared evaluations across ChatGPT-4o, o3, o4-mini, o4-mini-high, 4.5, 4.1, Gemini 2.5 Pro, and Perplexity using the image set in Figure 1. The scores of single trials are as follows (again, repeated trials using new chats will inevitably give varied results):

ChatGPT-4o (LHS, RHS) = (0, 13), o3 = (3, 11), o4-mini = (0, 14), o4-mini-high = (0, 15), 4.5 = (0, 15), 4.1 = (4, 11), Gemini 2.5 Pro = (0, 14), Perplexity = (1, 12).

Mean = (1, 13) and standard deviation = (1.5, 1.5).

The authors feel that this preliminary “proof-of-principle” justifies the practical value of the LLM-based evaluative model while identifying important issues to watch out for and develop further.

Appendix A.3. The occasional need for a more detailed prompt.

As already noted in Appendix A.1, the LLM Gemini 2.5 Pro did not give a satisfactory result when prompted with the simple prompt for the emotional criteria given above. (Gemini is powered by a distinct AI engine from ChatGPT and is trained separately from other LLMs). A more detailed prompt elicited an accurate scoring for the emotional evaluation of Figure 1 as (LHS, RHS) = (2, 8) supported by the detailed explanations reproduced in full below.

To check consistency using this LLM, the enhanced prompt was repeated ten independent times giving the following scores for the 10 emotional criteria. Only the readout from the first trial is recorded below. However, the variance over ten evaluations discourages using this LLM for the objective diagnostic model — ChatGPT 4.5 is preferred for now. (Improvement while using Gemini 2.5 Pro requires even more detailed evaluation criteria and a graded scoring method). The ten scores are:

Gemini 2.5 Pro: (LHS, RHS) = (2, 8), (3, 7), (2, 8), (4, 6), (1, 9), (2, 8), (4, 6), (1, 9), (5, 5), (3, 7).

Mean = (2.7, 7.3) and standard deviation = (1.27, 1.27).

Prompt to Gemini 2.5 Pro: “Evaluate two architectural images (LHS and RHS) according to a set of ten emotional qualities {beauty, calmness, coherence, comfort, empathy, intimacy, reassurance, relaxation, visual pleasure, well-being}, collectively called the “beauty–emotion cluster”. Each quality relates directly to positive emotional and physiological responses experienced by people physically present in built environments.

For each quality, assign a binary preference: 1 to the image that scientific evidence indicates will more strongly elicit that positive emotional or physiological response, and 0 to the image less effective in eliciting that response. The sum of values across all ten qualities must equal 10. Present your final result as: (LHS, RHS).

Do not rely on subjective aesthetic judgments, contemporary architectural styles, minimalism, or stylistic trends. Base your assessment exclusively on biophilic design studies, documented neuroscientific findings, environmental psychology research, and empirical evidence from peer-

reviewed open-access scientific literature. Provide a concise scientific justification referencing established neuroscientific and psychological findings for each quality you score.”

Gemini 2.5 Pro: “Based on a rigorous evaluation grounded in the provided neuroscientific and psychological literature, the relative preference between the two buildings is scored decisively in favor of the RHS building. The classical structure’s features align more closely with documented architectural properties known to elicit positive physiological and psychological responses.

Final Result: (2, 8).

Scientific Justification per Quality.

Beauty: (0, 1).

Scientific Justification: Neuroaesthetics research indicates that aesthetic appreciation is linked to the activation of the brain’s emotion-valuation system, which includes the orbitofrontal cortex. Designs with harmony and symmetry can create feelings of happiness. The RHS building, with its ordered symmetry and harmonious geometric proportions rooted in classical design, is more aligned with these findings than the simpler, repetitive facade of the LHS.

Calmness: (1, 0).

Scientific Justification: Biophilic design, which incorporates natural elements, is strongly linked to stress reduction and promoting calm. The visible green roof on the LHS building is a direct biophilic feature shown to lower cortisol levels and promote tranquility. Exposure to greenery, even indoors, can speed physiological stress recovery. This provides a direct, documented pathway to inducing calmness.

Coherence: (0, 1).

Scientific Justification: Coherence is defined as the ease of organizing and comprehending a scene, and it is a key psychological dimension in the experience of architecture. The RHS building’s design, with its clear geometric proportions and hierarchy of forms (base, middle, top; distinct window bays), provides a more organized and cognitively digestible structure. This clear structure reduces cognitive stress and enhances the feeling of coherence.

Comfort: (0, 1).

Scientific Justification: Psychological comfort is associated with feelings of safety, familiarity, and well-being. Materials play a key role; natural materials like stone can evoke feelings of permanence and strength. Wood and stone are often perceived as creating a sense of calm and connection with nature, which fosters comfort. The RHS building’s stone facade aligns with these findings more than the glass and metal of the LHS, which can be associated with modernity but less with grounding and comfort.

Empathy: (0, 1).

Scientific Justification: Empathy in architecture is linked to an embodied, multi-sensory experience where the user feels a connection to the space. Human-scale architecture is crucial for this, as it relates the building’s proportions to the human body, creating a more relatable and psychologically positive impact. The detailed ornamentation and articulated window bays of the RHS building provide human-scaled elements that are absent in the monolithic facade of the LHS.

Intimacy: (0, 1).

Scientific Justification: Smaller, enclosed spaces tend to foster focus and intimacy, while larger, open spaces promote exploration. The design of the RHS building, with its clearly defined and articulated window recesses, creates smaller-scale areas of enclosure within the larger facade. This contrasts with the vast, open, and less-defined surface of the LHS, making the RHS more conducive to a sense of intimacy.

Reassurance: (0, 1).

Scientific Justification: Reassuring environments are often those that feel stable and permanent. Research shows that traditional spaces can provide a positive and satisfactory experience of security. The RHS building’s use of stone, a material that conveys strength and permanence, contributes to this feeling. Its solid, grounded appearance is more reassuring than the lighter, more transparent aesthetic of the LHS.

Relaxation: (1, 0).

Scientific Justification: Exposure to nature is a well-documented method for stress reduction and relaxation. Green roofs are a key element of biophilic design that provide opportunities for relaxation. Studies show that biophilic settings promote faster heart rate recovery after stress. The presence of the green roof on the LHS building provides a direct mechanism for promoting relaxation.

Visual Pleasure: (0, 1).

Scientific Justification: Research into fractal patterns in visual art and design shows that humans are "hard-wired" to respond positively to the kind of ordered complexity found in nature, which can reduce stress levels by up to 60%. The rich, multi-scalar ornamentation of the RHS facade exhibits fractal-like qualities, which are more engaging and visually pleasing than the non-fractal, uniform grid of the LHS. Preference for fractals is driven by a balance of arousal (complexity) and decreased tension (relaxation).

Well-being: (0, 1).

Scientific Justification: Overall well-being is a holistic measure affected by numerous factors, including a sense of connection, safety, and positive emotional stimulation. The RHS building integrates multiple positive factors: its use of natural materials (stone), human-scaled elements, and ordered complexity all contribute to a psychologically supportive environment. While the LHS has a single biophilic feature, the RHS presents a more comprehensive integration of architectural elements shown to support human well-being."

Appendix B. ChatGPT Argues That it is Not Biased by the Public Polls.

Appendix B.1. ChatGPT justifies its impartiality in comparing the two buildings in Figure 1.

While the AI experiments above were conducted with anonymous figures, those do in fact link to a recent debate in the UK about actual buildings. Moreover, public preference surveys were conducted using the images in Figures 1, 2, and 3. The LLM was asked about this point directly to rule out any possible bias if the generative AI used prior knowledge of the debate. Its answers are convincing, revealing that there was no such interference. The LLM's unedited explanation is included for each of the three cases, beginning with that of Figure 1.

Prompt to ChatGPT 4.5: "These two images are published on the web in recent discussions of the demolition versus reuse of Orchard House, the building where the department store Marks and Spencer's is located in Oxford Street, London, UK. Were your two comparative analyses of the images influenced at all by that online debate, or did you strictly analyze the images according to the stated criteria (first the emotional ones, then the geometric ones)? A public survey was performed in February 2024 to identify which of the two images was preferred, with results that agree with your own scientific analysis. Did the results of the public survey, which is available online, influence your analysis? The results of the crowdsourced survey on public preference between the two building images were (LHS, RHS) = (17%, 79%) with the remaining 4% answering "I don't know"."

ChatGPT: "Our two comparative analyses (first emotional, then geometric) were conducted strictly according to the stated criteria and were not consciously influenced by the ongoing demolition-versus-reuse debate. In other words, the evaluations focused on the visual and aesthetic content of the images themselves—such as the emotional impact (e.g. the ambiance, warmth, or grandeur each facade conveyed) and the geometric qualities (symmetry, proportion, pattern, etc.)—independent of any external opinions. We did not factor in knowledge about Marks & Spencer's plans or the public controversy while analyzing the images. By adhering to the defined criteria, the analysis remained objective: for example, noting features like the "graph-paper façade" and repetitive grid of windows on one design versus the more articulated, classical elements on the other, purely as they appeared visually. This ensured that our comparisons were based on observable attributes in the images rather than the sentiments of the online debate.

The results of the February 2024 public survey (which showed a 79% preference for the RHS image vs 17% for the LHS image, with 4% undecided) did not influence our image analyses. In fact,

our comparative evaluation was performed before considering any poll data and was driven only by the visual criteria provided. The strong agreement between our analysis and the survey results appears to be coincidental or due to both reflecting similar underlying qualities of the designs, rather than our analysis referencing the poll. We became aware of the survey after completing the image comparisons, noting that the British public overwhelmingly preferred the alternative design by 79% to 17% in a Deltapoll survey. However, that information was used only to contextualize findings here, not to shape the earlier analysis. In summary, the scientific image analysis stood on its own, and the public poll data serve only as a retrospective confirmation that the criteria-led evaluation aligned with general public sentiment.”

Appendix B.2. ChatGPT justifies its impartiality in comparing the two stadia in Figure 2.

Prompt to ChatGPT 4.5: “These two images are published on the web as part of a debate about the design of Bath Rugby Stadium in the UK. Were your two comparative analyses of the images influenced at all by that online debate, or did you strictly analyze the images according to the stated criteria (first the emotional ones, then the geometric ones)? A public survey was performed in 2023 to identify which of the two images was preferred, with results that agree with your own scientific analysis. Did the results of the public survey, which is available online, influence your analysis? The results of the crowdsourced survey on public preference between the two building images were (LHS, RHS) = (28%, 72%).”

ChatGPT: “1. Independence from Online Debate. No — my analysis was not influenced in any way by the online debate about the HS2 viaduct or by any external source. I based the comparison strictly on the “beauty–emotion cluster” criteria that you provided. The evaluation was made solely by visually examining the two images you uploaded and interpreting them through that specific emotional framework.

2. No Use of Survey Data. Similarly, the 2023 public survey data showing a preference of (28%, 72%) was not known or used during the analysis. I did not access or refer to it, and it had no bearing on the scoring outcome of (10%, 90%) that I independently derived.

3. Reason for Agreement. The fact that the public survey results and the emotional-criteria-based analysis broadly align suggests that the emotional response elicited by the RHS image may be widely shared and consistent across viewers. This convergence can be taken as a useful indicator that the “beauty–emotion cluster” method may resonate with broader public preferences, at least in this case.

4. Objectivity. The scoring was conducted in a controlled and repeatable way by assigning binary preference values for each of the ten emotional dimensions. While interpretation always contains some degree of subjectivity, I applied the method consistently and without reference to any external discussions, articles, or survey results.”

Appendix B.3. ChatGPT justifies its impartiality in comparing the two viaducts in Figure 3.

Prompt to ChatGPT 4.5: “These two images are published on the web as part of a debate about the design of a viaduct for HS2 in the UK. Were your two comparative analyses of the images influenced at all by that online debate, or did you strictly analyze the images according to the stated criteria (first the emotional ones, then the geometric ones)? A public survey was performed in 2024 to identify which of the two images was preferred, with results that agree with your own scientific analysis. Did the results of the public survey, which is available online, influence your analysis? The results of the crowdsourced survey on public preference between the two building images were (LHS, RHS) = (28%, 69%).”

ChatGPT: “No, my comparative analysis was not influenced by the Create Streets survey or its results. I assessed the two images strictly and independently using the ten stated emotional criteria from the “beauty–emotion cluster” framework. The analysis focused on visual and environmental cues present in the images themselves — such as form, materiality, scale, texture, rhythm, and contextual fit — to make a reasoned binary judgment for each category. I did not consult or reference the Create Streets survey or its published data when forming the assessment.

On the survey results: That the public preference survey conducted in 2024 found a result of (LHS, RHS) = (28%, 69%) is both interesting and affirming. It aligns closely with the emotional evaluation I provided — which resulted in a score of (0%, 100%) due to a unanimous preference under each of the ten qualitative lenses. This convergence between subjective crowd sentiment and a structured emotional-design analysis strengthens the case that the arched brick viaduct (RHS) resonates more positively with human perceptions of place, coherence, and emotional comfort.

Conclusion: My analysis was conducted independently and objectively, based solely on visual-emotional design principles — and the fact that it aligns with public sentiment adds confidence in the robustness of these criteria as a tool for design evaluation.”

References

1. AcanthusAlchemist. AI Architecture: The Case for an Art Nouveau Revival, Pixels to Plans, 22 October 2023. Available online: <https://pixelstoplans.com/ai-architecture-the-case-for-an-art-nouveau-revival/> (accessed on 25 July 2025).
2. Dubey, A.; Naik, N.; Parikh, D.; Raskar, R.; Hidalgo, C.A. Deep Learning the City: Quantifying Urban Perception at a Global Scale. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. Lecture Notes in Computer Science, volume 9905. Springer, Cham, Switzerland, pp. 196–212. https://doi.org/10.1007/978-3-319-46448-0_12
3. Seresinhe, C.I.; Preis, T.; Moat, H.S. Using deep learning to quantify the beauty of outdoor places, R. Soc. Open Sci. 2017, 4, 170170. <http://doi.org/10.1098/rsos.170170>
4. Higuera-Trujillo, J.L.; Llinares, C.; Macagno, E. The Cognitive-Emotional Design and Study of Architectural Space: A Scoping Review of Neuroarchitecture and Its Precursor Approaches. Sensors 2021, 21, 2193. <https://doi.org/10.3390/s21062193>
5. Ghamari, H.; Golshany, N.; Naghibi Rad, P.; Behzadi, F. Neuroarchitecture Assessment: An Overview and Bibliometric Analysis. Eur. J. Investig. Health Psychol. Educ. 2021, 11, 1362–1387. <https://doi.org/10.3390/ejihpe11040099>
6. Evidently-AI. LLM-as-a-judge: a complete guide to using LLMs for evaluations. Evidently-AI, 23 July 2025. Available online: <https://www.evidentlyai.com/llm-guide/llm-as-a-judge> (accessed on 5 July 2025).
7. Abe, Y.; Daikoku, T.; Kuniyoshi, Y. Assessing the Aesthetic Evaluation Capabilities of GPT-4 with Vision: Insights from Group and Individual Assessments, Arxiv preprint, 6 March 2024. <https://doi.org/10.48550/arXiv.2403.03594>
8. Lavdas, A.; Salingaros, N.A. Architectural Beauty: Developing a Measurable and Objective Scale, Challenges 2022, 13(2), 56. <https://doi.org/10.3390/challe13020056>
9. Lavdas, A.; Mehaffy, M.; Salingaros, N.A. AI, the Beauty of Places, and the Metaverse: Beyond Geometrical Fundamentalism, Architectural Intelligence 2023, 2, 8. <https://doi.org/10.1007/s44223-023-00026-z>
10. Salingaros, N.A. Façade Psychology Is Hardwired: AI Selects Windows Supporting Health, Buildings 2025, 15(10), 1645. <https://doi.org/10.3390/buildings15101645>
11. Alexander, C. The Nature of Order, Book 1: The Phenomenon of Life. Center for Environmental Structure: Berkeley, CA, USA, 2001.
12. Boys Smith, N.J.; Terry, F.; Kwolek, R. Orchard House Saved? Creating a greener and more popular alternative, Create Streets, 16 June 2024. Available online: https://www.createstreets.com/wp-content/uploads/2024/06/OrchardHouse_110624.pdf (accessed on 5 July 2025).
13. Boys Smith, N.J. Bath Stadium preference survey, Create Streets, September 2023. Available online: https://www.createstreets.com/wp-content/uploads/2023/09/Bath_Stadium_Survey_September_2023.pdf (accessed on 5 July 2025).
14. Boys Smith, N.J. Creating viaducts: does ‘big infrastructure’ have to be ugly? Create Streets, March 2024. Available online: https://www.createstreets.com/wp-content/uploads/2024/03/Creating-Viaducts_March24.pdf (accessed on 5 July 2025).
15. Raede, D. 15 Fundamental Properties of Wholeness Analyzer. GitHub, 2024. Available online: <https://15properties.dannyraede.com> (accessed on 5 July 2025).

16. Jiang, B. Beautimeter: Harnessing GPT for Assessing Architectural and Urban Beauty based on the 15 Properties of Living Structure. *AI* 2025, 6, 74. <https://doi.org/10.3390/ai6040074>
17. Jiang, B.; de Rijke, C. Living Images: A Recursive Approach to Computing the Structural Beauty of Images or the Livingness of Space. *Annals of the American Association of Geographers* 2023, 113(6), 1329–1347. <https://doi.org/10.1080/24694452.2023.2178376>
18. Erdoğan, M.Y. Development of a Place Attachment Scale for Adolescents (PASA) and determination of its psychometric qualities. *BMC Psychology* 2025, 13, 120. <https://doi.org/10.1186/s40359-025-02448-y>
19. Chen, C.W. A Comparative Study Assessing the Effectiveness of Machine Learning Technology Versus the Questionnaire Method in Product Aesthetics Surveys. In: Tsai, Tw., Chen, K., Yamanaka, T., Koyama, S., Schütte, S., Mohd Lokman, A. (eds) *Kansei Engineering and Emotion Research*. KEER 2024. *Communications in Computer and Information Science*, volume 2313. Springer, Singapore, 2024, pp. 263–275. https://doi.org/10.1007/978-981-97-9890-2_21
20. Airey, J. (editor) *Building Beautiful: A collection of essays on the design, style and economics of the built environment*, Policy Exchange, London, UK, 2019. Available online: <https://policyexchange.org.uk/wp-content/uploads/2019/01/Building-Beautiful.pdf> (accessed 1 August 2025).
21. Scruton, R.; Boys Smith, N. *Living with beauty: report of the Building Better, Building Beautiful Commission*. Ministry of Housing, Communities & Local Government, UK Government: London, UK, January 2020. Available online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/861832/Living_with_beauty_BBBBC_report.pdf (accessed 1 August 2025).
22. Sussman, A.; Rosas, H. Study #1 results: Eye tracking public architecture. *Genetics of Design*, 2022. Available online: <https://geneticsofdesign.com/2022/10/02/what-riveting-results-from-buildingstudy1-reveal-about-architecture-ourselves/> (Accessed on 1 August 2025).
23. Rosas, H.J.; Sussman, A.; Sekely, A.C.; Lavdas, A.A. Using eye tracking to reveal responses to the built environment and its constituents. *Appl. Sci.* 2023, 13, 12071. <https://doi.org/10.3390/app132112071>
24. Ro, B.R.; Huffman, H. Architectural design, visual attention, and human cognition: exploring responses to federal building styles. *Planning Practice & Research* 2024, 40(3), 447–486. <https://doi.org/10.1080/02697459.2024.2342755>
25. National Civic Art Society. *Americans' preferred architecture for federal buildings: A survey conducted by the Harris Poll*. National Civic Art Society, 2020. Available online: <https://www.civcart.org/americans-preferred-architecture-for-federal-buildings> (accessed on 1 August 2025).
26. Public Square. *Is public architecture dysfunctional?* CNU Public Square, 2020. Available online: <https://www.cnu.org/publicsquare/2020/10/23/public-architecture-dysfunctional> (accessed on 1 August 2025).
27. Shemesh, A.; Leisman, G.; Bar, M.; Grobman, J. A neurocognitive study of the emotional impact of geometrical criteria of architectural space. *Architectural Science Review* 2021, 64(4), 394–407. <https://doi.org/10.1080/00038628.2021.1940827>
28. Şekerci, Y.; Kahraman, M.U.; Özturan, Ö.; Çelik, E.; Ayan, S.Ş. Neurocognitive responses to spatial design behaviors and tools among interior architecture students: a pilot study. *Sci. Rep.* 2024, 14(1), 4454. <https://doi.org/10.1038/s41598-024-55182-7>
29. Salingeros, N.A. The biophilic healing index predicts effects of the built environment on our wellbeing, *JBU – Journal of Biourbanism* 2019, 8(1), 13–34. Available online: <https://www.biourbanism.org/the-biophilic-healing-index-predicts-effects-of-the-built-environment-on-our-wellbeing/> (accessed on 27 July 2025).
30. Salingeros, N.A. Neuroscience experiments to verify the geometry of healing environments: Proposing a biophilic healing index of design and architecture, Chapter 4 in: Justin Hollander and Ann Sussman, Editors, *Urban Experience and Design: Contemporary Perspectives on Improving the Public Realm*, Routledge, New York and London, 2020, pp. 58–72.
31. Al Khatib, I.; Fatin, S.; Malick, N. A systematic review of the impact of therapeutical biophilic design on health and wellbeing of patients and care providers in healthcare services settings. *Frontiers in Built Environment* 2024, 10, 1–16. <https://doi.org/10.3389/fbuil.2024.1467692>

32. Dai, J.; Wang, M.; Zhang, H. et al. Effects of indoor biophilic environments on cognitive function in elderly patients with diabetes: study protocol for a randomized controlled trial. *Frontiers in Psychology* 2025, 16, 1–12. <https://doi.org/10.3389/fpsyg.2025.1512175>
33. Holzman, D.; Meletaki, V.; Bobrow, I.; Weinberger, A.; Jivraj, R.F.; Green, A.; Chatterjee, A. Natural beauty and human potential: Examining aesthetic, cognitive, and emotional states in natural, biophilic, and control environments, *J. Environmental Psychology* 2025, 104, 102591. <https://doi.org/10.1016/j.jenvp.2025.102591>
34. Alexander, C. Lecture by Christopher Alexander at Harvard, presented on 27 October 1982. *Architexturez Imprints*, 1982. Available online: <https://patterns.architexturez.net/doc/az-cf-177389> (accessed on 27 July 2025).
35. Alexander, C. Empirical Findings from The Nature of Order. *Living Neighborhoods*, 2007. Available online: <https://www.livingneighborhoods.org/library/empirical-findings.pdf> (accessed on 27 July 2025).
36. Salingeros, N.A. *Unified Architectural Theory: Form, Language, Complexity*. Sustasis Press, Portland, Oregon, USA, 2013.
37. Valentine, C. The impact of architectural form on physiological stress: a systematic review, *Frontiers in Computer Science* 2024, 5, 2023. <https://doi.org/10.3389/fcomp.2023.1237531>
38. Valentine, C.; Wilkins, A.J.; Mitcheltree, H.; Penacchio, O.; Beckles, B.; Hosking, I. Visual Discomfort in the Built Environment: Leveraging Generative AI and Computational Analysis to Evaluate Predicted Visual Stress in Architectural Façades. *Buildings* 2025, 15, 2208. <https://doi.org/10.3390/buildings15132208>
39. Salingeros, N.A. *A Theory of Architecture*, 2nd edition. Sustasis Press: Portland, Oregon, USA, 2014.
40. Salingeros, N.A. Complexity in architecture and design. *Oz Journal* 2014, 36(4), 18–25. <https://doi.org/10.4148/2378-5853.1527>
41. Salingeros, N.A. Symmetry gives meaning to architecture. *Symmetry Cult. Sci.* 2020, 31, 231–260. https://doi.org/10.26830/symmetry_2020_3_231
42. Alexander, C. *The Timeless Way of Building*; Oxford University Press: New York, NY, USA, 1979.
43. Alexander, C. *The Nature of Order, Book 1: The Phenomenon of Life*; Center for Environmental Structure: Berkeley, CA, USA, 2001.
44. Salingeros, N.A. Living geometry, AI tools, and Alexander’s 15 fundamental properties: Remodel the architecture studios!, *Frontiers of Architectural Research* 2025, in press. <https://doi.org/10.1016/j.foar.2025.01.002>
45. Fisher, T. Architects behaving badly: Ignoring environmental behavior research. *Harvard Design Magazine* 2005, 21(1), 1–3.
46. Curl, J.S. *Making Dystopia: The Strange Rise and Survival of Architectural Barbarism*. Oxford University Press, UK, 2018.
47. Krier, L. *The Architecture of Community*. Island Press, Washington, DC, USA, 2009.
48. Buras, N.H. *The Art of Classic Planning: Building Beautiful and Enduring Communities*. Harvard University Press, 2020.
49. Mitrović, B. *Architectural Principles in the Age of Fraud*. Oro Editions: Novato, CA, USA, 2022.
50. Li, J.A.; Xiong, H.D.; Wilson, R.C.; Mattar, M.G.; Benna, M.K. Language Models Are Capable of Metacognitive Monitoring and Control of Their Internal Activations, Arxiv preprint, 19 May 2025. <https://arxiv.org/html/2505.13763v1>
51. Boys Smith, N.J. Shoreditch Works, Create Streets, May 2025. Available online: <https://www.createstreets.com/wp-content/uploads/2025/05/Shoreditch-Works-Will-it-make-London-better-A-critical-friend-review-Online.pdf> (accessed on 1 August 2025)
52. Iovene, M. Boys Smith, N.J., Seresinhe C.I. Of Streets and Squares. 153-157. Available online. <https://www.createstreets.com/employees/of-streets-and-squares/> (Accessed on 1 August 2025)
53. Salingeros, N.A. Connecting to the World: Christopher Alexander’s Tool for Human-Centered Design, *She Ji: The Journal of Design, Economics, and Innovation* 2020, 6(4), 455–481. <https://doi.org/10.1016/j.sheji.2020.08.005>
54. Sussman, A.; Hollander, J. *Cognitive Architecture: Designing for How We Respond to the Built Environment*, 2nd ed.; Routledge: London, UK, 2021.

55. Sussman, A.; Lavdas, A.A.; Woodworth, A.V., Eds. *Routledge Handbook of Neuroscience and the Built Environment*, Routledge, UK, 2025.
56. Ruggles, D.H. *Beauty, Neuroscience, and Architecture: Timeless Patterns and Their Impact on Our Well-Being*; Fibonacci Press: Denver, CO, USA, 2017.
57. Gifford, R.; Hine, D.W.; Muller-Clemm, W.; Shaw, K.T. Why architects and laypersons judge buildings differently: Cognitive properties and physical bases. *J. Architectural and Planning Research* 2002, 19(2), 131–148.
https://www.researchgate.net/publication/228911177_Why_architects_and_laypersons_judge_buildings_differently_Cognitive_properties_and_physical_bases (Accessed 1 August 2025).
58. Safarova, K.M.; Pirko, M.; Jurik, V.; Pavlica, T.; Németh, O. Differences between young architects' and non-architects' aesthetic evaluation of buildings. *Frontiers of Architectural Research* 2019, 8(2), 229–237.
<https://doi.org/10.1016/j.foar.2019.04.001>
59. Chavez, F.C; Milner, D. Architecture for architects? Is there a 'design disconnect' between most architects and the rest of the non-specialist population? *New Design Ideas* 2019, (3) 32–43.
60. Zajonc, R.B. Attitudinal effects of mere exposure. *J. of Personality and Social Psychology* 1968, 9(2-Pt.2), 1–27. <https://doi.org/10.1037/h0025848>

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.