èkphrasis

46° Convegno Internazionale
dei Docenti delle Discipline della Rappresentazione
Congresso della Unione Italiana per il Disegno

Allucinazione eidomatica degli ordini architettonici nell'era delle Reti Neurali

Giulia Flenghi
Michele Russo
Luca J. Senatore

Abstract

L'uso delle Reti Neurali per la costruzione di contenuti visuali è oggi un tema che pone sfide di natura multi-disciplinare. La generazione automatica di immagini (*text-to-image*) e video (*text-to-video*) coinvolge molteplici ambiti applicativi ed introduce il problema della affidabilità dei contenuti. All'interno di questo filone di ricerca ricade il tema delle 'allucinazioni digital', intese come processo di inserimento non casuale di errori o dettagli inaspettati nelle immagini digitali. La presenza consistente di questi artefatti visivi all'interno delle immagini generate automaticamente dagli algoritmi di Deep Learning rappresenta un problema nell'utilizzo di questi strumenti, che mostrano grandi potenzialità ma altrettanti limiti nella costruzione di contenuti scientificamente affidabili. La ricerca, partendo da uno stato dell'arte sul tema, propone la sperimentazione comparativa fra differenti piattaforme *text-to-image* per la generazione di immagini che contengono elementi architettonici codificati dalla trattatistica e relativi all'ordine dorico. Suggerendo una prima classificazione delle allucinazioni visive, la ricerca mostra come questi strumenti non siano ancora in grado di proporre contenuti scientificamente corretti nell'ambito della architettura, aprendo interessanti scenari nella relazione fra il testo e le immagini.

Parole chiave
Allucinazioni digitali, *Text-to-image generation*, *Large Vision-Language Models* (LVLMs), affidabilità epistemologica, rappresentazione architettonica computazionale.

Da sinistra a destra: immagini generate utilizzando le piattaforme *DALL·E*, *Stable Diffusion* e *Midjourney*. La composizione evidenzia diverse interpretazioni della stessa tematica attraverso i differenti algoritmi di intelligenza artificiale (elaborazione a cura degli autori).

doi.org/10.3280/oa-1430-c900

Le allucinazioni digitali

L'uso delle Reti Neurali per la costruzione di contenuti visuali è oggi un tema che pone sfide di natura multi-disciplinare. La generazione automatica di immagini (*text-to-image*) e video (*text-to-video*) coinvolge molteplici ambiti applicativi ed introduce il problema della affidabilità dei contenuti. All'interno di questo filone di ricerca ricade il tema delle 'allucinazioni digitali', intese come processo di inserimento non casuale di errori o dettagli inaspettati nelle immagini digitali.
Per comprenderne le ragioni, è essenziale riportare in sintesi il processo che parte da descrizioni testuali (*input*) e arriva alla generazione delle immagini (*output*). Una Rete Neurale si basa su un modello di *Deep Learning* che prende decisioni in modo simile al cervello umano, utilizzando una struttura di nodi distribuiti su un numero di livelli variabile e che simula il funzionamento dei neuroni umani. Ogni nodo è connesso ad altri ed è caratterizzato da un peso e una soglia di attivazione. Se tale soglia viene superata a livello locale, il nodo invia dati al livello successivo, altrimenti il processo termina. Le Reti Neurali si basano su dati di addestramento per imparare e migliorare la loro precisione nel tempo, individuando specifici valori di soglia.
Quando la Rete Neurale riceve un *input* testuale, genera un'immagine basandosi sui pattern statistici appresi attraverso l'elaborazione di grandi quantità di immagini contenute in dataset prestabili, utili per l'apprendimento dell'algoritmo e implementabili nel tempo. In questo senso la qualità del dataset di addestramento è cruciale: maggiore è la sua dimensione, variabilità e corretta classificazione dei dati al suo interno, maggiore sarà la probabilità che l'algoritmo risponda correttamente alla richiesta. Tuttavia il modello, pur mostrando una incredibile velocità in termini di risposta alla domanda, non 'comprende' il senso profondo di ciò che rappresenta. Da questo deriva una diretta dipendenza dal dato di addestramento e alla *query*, non avendo la capacità di generalizzare la richiesta creando inferenze esterne.
Per questo motivo, se la richiesta testuale richiama delle informazioni che non sono contenute nei dati di addestramento oppure non è abbastanza specifica e presenta termini interpretabili, la macchina genera immagini o schematizzazioni errate e prive di senso [Del Campo, Leach 2022]. Questi risultati vengono codificati con il nome di 'allucinazioni digitali', nelle quali posso comparire informazioni errate, paradossali, surreali. Anche se tale termine è connesso al funzionamento del cervello umano, la risposta data dalla macchina a livello di comportamento è assimilabile a tale processo, portando a conseguenze imprevedibili.

Stato dell'arte

L'evoluzione dei modelli generativi di *Deep Learning* degli ultimi anni ha evidenziato significativi progressi nella generazione automatica di testi, immagini e video, migliorando notevolmente sia i tempi di generazione che la realisticità del risultato finale. Ma questa evoluzione ha alimentato anche il tema delle allucinazioni digitali, particolarmente rilevante nei *Large Language Models* (LLMs) e nei *Large Vision-Language Models* (LVLMs), influenzando la qualità delle risposte testuali o visive [Ji *et al.* 2023]. A livello scientifico si sottolineano alcune differenze nella definizione di *AI hallucination* in funzione dell'ambito applicativo. Alcuni studi propongono termini alternativi, mentre altri evidenziano la necessità di una standardizzazione della definizione, per facilitare lo sviluppo di strategie di mitigazione efficaci [Maleki Padmanabhan, Dutta 2024].
Le allucinazioni digitali possono derivare da diversi fattori, tra cui overfitting, distorsione nei dati di addestramento e complessità del modello. Un dataset non rappresentativo o contenente *bias* può portare a errori sistematici nell'*output*. I modelli multimodali, che combinano *input* testuali e visivi, soffrono in particolare di *Visual Hallucination* (VH), ossia errori nell'interpretazione delle immagini in compiti di *Visual Question Answering* (VQA). Questi errori sono amplificati dall'uso di immagini generate artificialmente (AIGC), che possono alterare le rappresentazioni interne del modello e aumentare la probabilità di identificare dettagli inesistenti. I VHTest, fra cui si annoverano *GPT-4V*, *LLaVA-1.5* e *MiniGPT-v2*, sono stati sviluppati per generare *benchmark* specifici e analizzare i modelli neurali. I risultati mostrano che il *fine-tuning* su dataset mirati può ridurre il tasso di allucinazione senza compromettere le

prestazioni generali [Huang *et al.* 2024]. Inoltre si evidenzia la presenza di un bias sistematico nelle allucinazioni causate da immagini sintetiche nei LVLMs, Tali immagini possono influenzare la rappresentazione interna dei modelli, portando a errori amplificati nei processi di elaborazione visiva [Gao *et al.* 2024].
La ricerca è in continua evoluzione, data la rapida trasformazione degli strumenti e la loro pervasività applicativa, per sviluppare metodi di controllo in grado di riconoscere e correggere automaticamente le discrepanze tra *input* e *output*. La finalità è quella di aumentare l'affidabilità dei modelli e implementare i meccanismi interni delle Reti Neurali generative. Per questo, la maggior parte delle ricerche è oggi rivolta alla definizione di strategie volte a migliorare la coerenza semantica e ridurre gli errori generativi [Ramesh *et al.* 2021; Tonmoy *et al.* 2024]. Numerosi studi evidenziano come la qualità degli *output* dipenda fortemente dalla quantità, varietà e qualità dei dati di addestramento, oltre che dall'ottimizzazione dei parametri del modello [Rombach *et al.* 2021]. Inoltre, l'introduzione di filtri e soglie probabilistiche può migliorare la coerenza e l'accuratezza dei risultati [Biten, Gómez, Karatzas 2022]. Approcci innovativi includono l'adozione di architetture ibride che combinano tecniche di trasformatori con modelli di diffusione, impiegando strategie di *fine-tuning* volte a mitigare l'insorgenza di artefatti indesiderati. Alcuni ricercatori interpretano tali allucinazioni come una potenziale fonte di creatività, capace di suggerire nuove prospettive artistiche e progettuali [Goodfellow *et al.* 2014].

Un confronto fra le piattaforme

Nella fase sperimentale sono state testate quattro piattaforme: *DALL-E 3* (tramite *Chat GPT*), *Imagen 3* (tramite *Gemini*), *Midjourney* e *Stable Diffusion* (*stand alon*e e integrato con *Control net*). L'obiettivo di questa fase è stato quello di verificare la capacità dei diversi sistemi di generare immagini dell'ordine dorico, partendo dai testi di Vitruvio e successivamente integrando elementi visivi. L'approccio iniziale si è basato esclusivamente su *input* testuali, traducendo estratti di Vitruvio dal volgare all'italiano contemporaneo e all'inglese. I prompt sono stati affinati progressivamente: partendo da traduzioni automatiche, si è passati a descrizioni generali e in seguito di dettaglio su colonne, capitelli e facciate, integrandole con liste di parole chiave. Infine, è stata introdotta la lingua latina per verificarne il ruolo nel miglioramento dell'accuratezza delle immagini. Successivamente, sono stati aggiunti elementi visivi per affinare la risposta della Rete Neurale, introducendo illustrazioni tratte da Serlio, descrizioni automatiche generate da AI e fotografie di elementi reali. Dalle sperimentazioni condotte è emerso che i modelli non posseggono né i termini descrittivi dell'ordine dorico né immagini classificate, indipendentemente dalla lingua utilizzata nei *prompt*. Non avendo riferimenti precisi, i sistemi generano immagini errate. Alla richiesta di un 'capitello dorico', il sistema produce spesso capitelli ionici, poiché il termine 'capitello' è più frequentemente associato a quest'ultimo stile nei dataset di addestramento. Oppure, come nel caso di *DALL-E 3*, *Chat GPT* possiede al suo interno la definizione di ordine dorico, ma *DALL-E 3* produce comunque immagini errate a causa della mancanza di tag appropriati nei suoi dati di addestramento (fig. 1).

Fig. 1. Immagini generate con *DALL·E 3*. I *prompt* utilizzati, da sinistra a destra, sono: una descrizione dettagliata di un capitello dorico generata da *ChatGPT*; '*Doric column*'; un estratto dal testo di Vitruvio con la descrizione di un tempio dorico (elaborazione a cura degli autori).

Fig. 2. Immagini generate con *Imagen 3*. Da sinistra a destra, i *prompt* utilizzati sono: 'proporzionalmente capitello dorico', 'ordine dorico' e 'tempio dorico' (elaborazione a cura degli autori).

Fig. 3. Immagini generate con *MidJourney*. I *prompt* utilizzati, da sinistra a destra, sono: 'Colonna dorica', 'Capitello dorico', 'Base attica' (tipica dell'ordine ionico e corinzio) (elaborazione a cura degli autori).

Fig. 4. Immagine generata con *MidJourney*. Il *prompt* utilizzato è una descrizione del fronte di un tempio dorico ottenuta tramite la funzione */describe* (elaborazione a cura degli autori).

Fig. 5. Immagini generate con *Stable Diffusion*. Da sinistra a destra, i *prompt* utilizzati sono: 'echino', 'entasi', '*front view of an ancient Greek temple, with two columns on each side and one door in the middle. Doric order*' (elaborazione a cura degli autori).

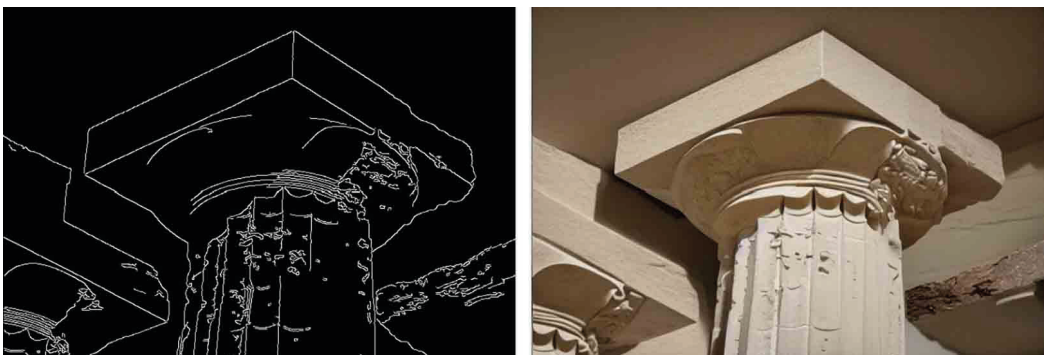

Fig. 6. Immagini generate con *Stable Diffusion* e *ControlNet* tramite il *prompt* 'Capitello dorico' e un'immagine di riferimento. A sinistra: la rilevazione dei contorni con *Canny edges* di *ControlNet*; a destra: l'immagine generata (elaborazione a cura degli autori).

Analogo comportamento è stato riscontrato con *Imagen 3*, utilizzato da *Gemini*, dove il chatbot possiede il concetto di ordine dorico, ma il generatore di immagini, non disponendo di riferimenti adeguati per riprodurlo correttamente, può dare luogo ad allucinazioni (fig. 2).
Al contrario, *MidJourney* offre un maggiore controllo nella generazione, grazie alla possibilità di combinare testo e immagini. Ciò migliora la qualità delle immagini rispetto agli altri strumenti, senza tuttavia eliminare del tutto le allucinazioni (fig. 3).
Alcuni risultati più accurati sono stati ottenuti con la funzione '*/describe*' e con *prompt* meno specifici, ma la mancanza di dati di addestramento adeguati continua a causare distorsioni nell'*output* (fig. 4).
Infine, anche *Stable Diffusion* presenta problemi simili: non avendo riferimenti precisi sull'ordine dorico, produce immagini approssimative. L'unico risultato parzialmente corretto è stato ottenuto generando la facciata di un tempio dorico, sebbene con errori evidenti (fig. 5).
L'integrazione con ControlNet ha consentito di migliorare l'accuratezza delle immagini generate, riducendo di conseguenza le allucinazioni. In questo caso sistema risulta meno influenzato dai dati di addestramento (errati o lacunosi) e si basa maggiormente sulla replica dei contorni di un'immagine preesistente attraverso la funzione *Canny edge*. Questo introduce dei paletti nella definizione della forma, ma non implica una corretta rappresentazione dell'ordine dorico (fig. 6).
In generale, tutte le piattaforme testate hanno evidenziato la mancanza di dati di addestramento specifici sull'ordine dorico, sia di natura testuale che visiva. Ma la implementazione di tali dataset pone il problema di accessibilità dei dati delle Reti Neurali 'commerciali'.

Analisi critica

Lo studio fin qui condotto ha evidenziato come la Rete Neurale, costruita per simulare i processi logici propri del ragionamento umano attraverso inferenze, in mancanza di adeguate sollecitazioni e informazioni di partenza possa produrre risultati incoerenti rispetto alle richieste. Dalla sperimentazione è emersa l'importanza nella definizione di descrizioni testuali corrette, condizione necessaria ma non sufficiente per l'eliminazione delle allucinazioni. Que-

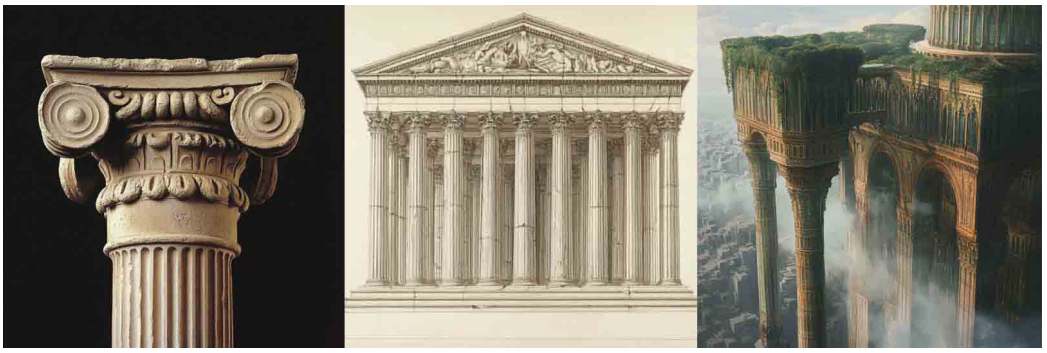

Fig. 7. Tre livelli di allucinazione generativa. Da sinistra a destra: lieve, intermedia e complessa (elaborazione a cura degli autori).

sta mancata di *consecutio* fra *input* e *output*, mediata da un modello esterno con specifiche funzionalità, dimostra come ad oggi tali strumenti non siano affidabili. Questa variabilità del risultato può infatti risultare interessante nell'ambito della progettazione, nel quale un processo creativo può ricevere stimoli positivi dalla costruzione di inferenze inaspettate. Ma nel dominio della rappresentazione architettonica è essenziale generare contenuti formalmente corretti e di valore scientifico, altrimenti vi è il rischio di errate interpretazioni del reale e la trasmissione di una conoscenza sbagliata.
Paradossalmente, mentre i modelli LLM risultano molto affidabili, la mancata relazione diretta con i modelli LVLM porta ad una notevole frequenza delle allucinazioni. Se si considera questo fenomeno solamente come errore, possono essere applicate delle metriche per quantificare la percentuale di risultati corretti o errati rispetto a una griglia di riferimento. Tuttavia, quando manca una griglia di partenza, diventa più complesso determinare il livello di errore. Nel caso delle rappresentazioni architettoniche, è difficile stabilire metriche precise. Tuttavia, si possono identificare genericamente tre livelli di allucinazione (fig. 7):
1. lieve: generata nei livelli meno profondi della Rete, con dati di *input* parzialmente presenti. Si manifesta con piccoli errori, come dettagli incoerenti, discontinuità materiche minime o imprecisioni prospettiche. Questi errori sono visibili solo da osservatori attenti ed esperti;
2. intermedia: derivata da una minore presenza di dati di *input* e una maggiore profondità di elaborazione della Rete. Introduce errori evidenti come l'assenza di elementi architettonici, la presenza di forme prive di senso strutturale e prospettive distorte. Questa allucinazione può causare una errata lettura del risultato. Un utente senza conoscenze specifiche potrebbe ancora percepire il contenuto come coerente;
3. complessa: generata nei livelli più profondi della rete con pochi o inesistenti dati di *input*. Produce contenuti irreali, con assenza di intere porzioni architettoniche, gravi discontinuità formali e contesti completamente incoerenti, facilmente riconoscibili anche da utenti inesperti.

Conclusioni e prospettive

La ricerca mostrata ha avuto come obbiettivo quello di verificare le attuali capacità di alcune Reti Neurali generaliste text-to-image di produrre immagini scientificamente corrette, focalizzandosi nel dominio della rappresentazione dell'ordine dorico. Lo studio ha evidenziato come, allo stato attuale, tali reti neurali non siano ancora in grado di generare immagini scientificamente attendibili, producendo artefatti chiamati 'allucinazioni digitali'. Per ridurre tale fenomeno, è fondamentale utilizzare dataset bilanciati e ben strutturati, implementando modelli specifici e fasi di testing continue nelle quali la supervisione umana è essenziale per filtrare, correggere e validare i risultati. Questi aspetti evidenziano l'esigenza di mettere a sistema differenti competenze per arrivare ad un risultato significativo.
I risultati ottenuti dalla sperimentazione hanno permesso di tracciare alcuni possibili scenari futuri della ricerca. Nel breve termine ci si può concentrare sul miglioramento del linguaggio e dei riferimenti testuali, introducendo quei termini tecnici che possono favorire la ricerca delle poche illustrazioni scientificamente corrette presenti nei database. Inoltre, l'introdu-

zione delle traduzioni originali dei trattati in latino e greco può favorire la costruzione di inferenze. Ma la mancanza di un dato correttamente classificato rimane il limite principale. Per questo, una seconda fase della ricerca deve concentrarsi sulla definizione di strategie per l'addestramento mirato delle Reti Neurali generaliste. Questo può avvenire attraverso una implementazione dei dataset con immagini correttamente classificate e/o l'introduzione di filtri in grado di introdurre dei paletti nella ricerca, migliorando il *fine-tuning* del risultato. Tale procedura ha il limite principale nella limitata accessibilità dei dati di queste reti. Un'ultima strada consiste nella generazione di un nuovo modello specializzato nella generazione di immagini coerenti e scientificamente attendibili, progettando una fase di addestramento con immagini correttamente classificate inserite in un database multilivello in grado di relazionarsi con tutte le possibili richieste in ambito architettonico. Questa strada, seppure onerosa in termini di risorse informatiche e tempo, risulta attualmente l'unica in grado di costruire un prodotto in grado di ampliare, attraverso l'utilizzo di questa tecnologia, le possibilità di esplorazione offerte a studiosi nell'ambito dell'architettura.

Riconoscimenti

La stesura dei paragrafi è stata organizzata come segue: Giulia Flenghi ha curato *Stato dell'arte* e *Un confronto fra le piattaforme*; Michele Russo ha sviluppato il paragrafo *Analisi critica*; Luca J. Senatore si è occupato di *Le allucinazioni digitali* e *Conclusioni* e *prospettive*.

Riferimenti bibliografici

Biten, A. F., Gómez, L., Karatzas, D. (2022). Let there be a clock on the beach: Reducing object hallucination in image captioning. In *IEEE/CVF Winter Conference on Applications of Computer Vision* (WACV), Waikoloa, pp. 2473-2482. https://doi.org/10.1109/WACV51458.2022.00253.

Del Campo, M., Leach, N. (Guest Eds.). (2022). *Architectural Design*, *Machine Hallucinations: Architecture and Artificial Intelligence,* Vo. 92, 3.

Gao, Y., Wang, J., Lin, Z., Sang, J. (2024). *AIGCs confuse AI too: Investigating and explaining synthetic image-induced hallucinations in large vision-language models.* arXiv preprint arXiv:2403.08542. https://arxiv.org/abs/2403.08542.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. (2014). *Generative adversarial networks. Advances in Neural Information Processing Systems*, 27.

Huang, W., Liu, H., Guo, M., Gong, N. Z. (2024). *Visual hallucinations of multi-modal large language models.* arXiv preprint arXiv:2402.14683. https://arxiv.org/abs/2402.14683.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Xu,Y., Ishii, E., Jin Bang, Y., Madotto, A., Fung, P. (2023). Survey of hallucination in natural language generation. In *ACM Computing Surveys,* 55(12), Article 248, pp. 1-38. https://doi.org/10.1145/3571730.

Maleki, N., Padmanabhan, B., Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. In *IEEE Conference on Artificial Intelligence* (CAI), Singapore, pp. 133-138. https://doi.org/10.1109/CAI59869.2024.00033.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A. Chen, M., Sutskever, I. (2021). *Zero-Shot Text-to-Image Generation*. arXiv:2102.12092 https://doi.org/10.48550/arXiv.2102.12092.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2021). *High-resolution image synthesis with latent diffusion models.* arXiv preprint arXiv:2112.10752. https://arxiv.org/abs/2112.10752.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., Das, A. (2024). *A comprehensive survey of hallucination mitigation techniques in large language models*. arXiv preprint arXiv:2401.01313. https://arxiv.org/abs/2401.01313.

Autori
Giulia Flenghi, Sapienza Università di Roma, giulia.flenghi@uniroma1.it
Michele Russo, Sapienza Università di Roma, m.russo@uniroma1.it
Luca J. Senatore, Sapienza Università di Roma, luca.senatore@uniroma1.it

Per citare questo capitolo: Giulia Flenghi, Michele Russo, Luca J. Senatore (2025). Allucinazione eidomatica degli ordini architettonici nell'era delle Reti Neurali. In L. Carlevaris *et al.* (a cura di). *èkphrasis. Descrizioni nello spazio della rappresentazione/èkphrasis. Descriptions in the space of representation.* Atti del 46° Convegno Internazionale dei Docenti delle Discipline della Rappresentazione. Milano: FrancoAngeli, pp. 2777-2792. DOI: 10.3280/oa-1430-c900.

ISBN 9788835182412

ekphrasis

46th International Conference
of Representation Disciplines Teachers
Congress of Unione Italiana per il Disegno

Eidomatic Hallucination of Architectural Orders in the Age of Neural Networks

Giulia Flenghi
Michele Russo
Luca J. Senatore

Abstract

Neural Network application for visual content generation is a hot topic with multidisciplinary challenges. The automatic generation of images (text-to-image) and videos (text-to-video) involves multiple application domains and introduces the problem of content reliability. Within this line of research falls the topic of 'digital hallucinations,' the process of non-random inclusion of errors or unexpected details in digital images. The consistent presence of these visual artefacts in images automatically generated by Deep Learning algorithms defines a problem of platform usability. They show great potential but many limitations in constructing scientifically reliable content. The research proposes comparative experimentation between different text-to-image platforms for generating images containing architectural elements encoded by the treatise and related to the Doric order. In particular, starting from an initial classification of visual hallucinations, the research shows how these tools cannot propose scientifically correct content in architecture, opening up engaging scenarios in the relationship between text and images.

Keywords
Digital hallucinations, Text-to-image generation, Large Vision-Language Models (LVLMs), Epistemological reliability, Computational architectural representation.

Images generated using, from left to right, the platforms *DALL·E*, *Stable Diffusion*, and *Midjourney*. The composition highlights different interpretations of the same theme through various artificial intelligence algorithms (processing by the authors).

doi.org/10.3280/oa-1430-c900

Digital hallucinations

The use of Neural Networks to generate visual content poses multidisciplinary challenges. The automatic generation of images (text-to-image) and videos (text-to-video) involves multiple application domains and introduces the problem of content reliability. Within this line of research falls the topic of 'digital hallucinations,' understood as the process of non-randomly inserting errors or unexpected details into digital images.
To understand the origins of these hallucinations, it is important to briefly report the process that starts with textual descriptions (input) and arrives at image generation (output). A Neural Network is based on a Deep Learning model that makes decisions similar to the human brain's. It uses a structure of nodes distributed over a variable number of levels and simulates the functioning of human neurons. Each node is connected to others and has a weight and an activation threshold. If this threshold is exceeded locally, the node sends data to the next layer; otherwise, the process ends. Neural Networks rely on training data to learn and improve their accuracy over time by identifying specific threshold values.
When the Neural Network receives textual input, it generates an image based on statistical patterns learned by processing large quantities of images in predefined datasets, which helps train the algorithm and is implementable over time. In this sense, the quality of the training dataset is crucial: the more significant its size, variability and correct classification of the data, the greater the possibility that the algorithm will respond correctly to the query. However, while showing incredible speed in answering the query, the model does not 'understand' the meaning of what it represents. It follows a direct dependence on the training dataset and the query, not having the ability to generalise the query by creating external inferences.
For this reason, if the textual prompt invokes information that is not contained in the training data or is not specific enough and has interpretable terms, the machine generates erroneous and meaningless images or schematisations [Del Campo, Leach 2022]. These results are coded as 'digital hallucinations,' in which erroneous, paradoxical, surreal information can appear. Although this term is related to the functioning of the human brain, the response given by the machine at the level of behaviour can be likened to this process, leading to unpredictable consequences.

State of the Art

The evolution of Deep Learning generative models in recent years has shown significant advances in the automatic generation of text, images and video, significantly improving both the generation time and the realistic nature of the final result. However, this evolution has also fuelled the issue of digital hallucinations, which is particularly relevant in Large Language Models (LLMs) and Large Vision-Language Models (LVLMs), affecting the quality of textual or visual responses [Ji *et al.* 2023]. Some differences in the definition of AI hallucination depending on the application domain are highlighted at the scientific level. Different research proposes alternative terms, while others highlight the need to standardise the definition to facilitate the development of effective mitigation strategies [Maleki Padmanabhan, Dutta 2024].
Digital hallucinations can result from several factors, including overfitting, bias in the training data, and model complexity. A dataset that is unrepresentative or contains bias can lead to systematic errors in the output. Multimodal models, which combine textual and visual inputs, particularly suffer from Visual Hallucination (VH), i.e., errors in image interpretation in Visual Question Answering (VQA) tasks. These errors are amplified by artificially generated images (AIGCs), which can alter internal model representations and increase the likelihood of identifying nonexistent details. VHTests, including GPT-4V, LLaVA-1.5 and MiniGPT-v2, were developed to generate specific benchmarks and analyse neural models. Results show that fine-tuning targeted datasets can reduce the hallucination rate without compromising overall performance [Huang *et al.* 2024]. We also show evidence of systematic bias in hallucinations caused by synthetic images in LVLMs. Such images can affect the internal representation of models, leading to amplified errors in visual processing [Gao *et al.* 2024].

Given the rapid transformation of tools and their pervasiveness in application, research is constantly evolving to develop control methods that can automatically recognise and correct discrepancies between input and output. The aim is to increase the reliability of models and implement the internal mechanisms of Generative Neural Networks. Therefore, most research today focuses on defining strategies to improve semantic consistency and reduce generative errors [Ramesh *et al.* 2021; Tonmoy *et al.* 2024]. Numerous studies point out that the quality of outputs depends on the quantity, variety and quality of training data and the optimisation of model parameters [Rombach *et al.* 2021]. In addition, introducing probabilistic filters and thresholds can improve the consistency and accuracy of the results [Biten, Gómez, Karatzas 2022]. Innovative approaches include the adoption of hybrid architectures that combine transform techniques with diffusion models, employing fine-tuning strategies aimed at mitigating the occurrence of unwanted artefacts. Some researchers interpret such hallucinations as a potential source of creativity, capable of suggesting new artistic and design perspectives [Goodfellow *et al.* 2014].

A platform comparison

Four platforms were tested in the experimental phase: *DALL-E 3* (via Chat GPT), *Imagen 3* (via *Gemini*), *Midjourney*, and *Stable Diffusion* (stand-alone and integrated with Control Net). The objective of this phase was to test the ability of the different systems to generate images of the Doric order, starting with Vitruvius' texts and then integrating visual elements. The initial approach was based solely on textual input, translating excerpts from Vitruvius from vernacular to contemporary Italian and English. The prompts were progressively refined: starting with automatic translations, general and later detailed descriptions of columns, capitals, and facades were added, supplemented with lists of keywords. Finally, Latin was introduced to test its role in improving the accuracy of images. Next, visual elements were added to refine the response of the Neural Network by introducing illustrations from Serlio, automatic AI-generated descriptions, and photographs of real elements.
Experiments have shown that the models possess neither the descriptive terms of the Doric order nor classified images, regardless of the language used in the prompts. If the systems do not have precise references, they generate incorrect images. When asked for a 'Doric capital,' the system often produces Ionic capitals since 'capital' is most frequently associated with the latter style in the training datasets. In the case of *DALL-E 3*, *Chat GPT* has the definition of Doric order internally, but *DALL-E 3* still produces erroneous images due to the lack of appropriate tags in its training data (fig. 1).
Similar behaviour was found with *Imagen 3* integrated with *Gemini*. The chatbot owns the concept of Doric order, but the image generator's lack of adequate references to reproduce it correctly can result in hallucinations (fig. 2).
In contrast, *MidJourney* offers more control in the generation due to its ability to combine text and images. It improves image quality compared to the other tools but does not eliminate hallucinations (fig. 3).

Fig. 1. Images generated with *DALL·E 3*. From left to right, the prompts used, are: a detailed description of a Doric capital generated by *ChatGPT*; 'Doric column'; an excerpt from Vitruvius' text describing a Doric temple (processing by the authors).

Fig. 2. Images generated with *Imagen 3*. From left to right, the prompts used are: 'proportionally Doric capital,' 'Doric order,' and 'Doric temple' (processing by the authors).

Fig. 3. Images generated with *MidJourney*. From left to right, the prompts used are: 'Doric column,' 'Doric capital,' 'Attic base' (typical of the Ionic and Corinthian orders) (processing by the authors).

Fig. 4. Image generated with *MidJourney*. The prompt used is a description of the front of a Doric temple obtained through the /describe function (processing by the authors).

Fig. 5. Images generated with *Stable Diffusion*. From left to right, the prompts used are: 'echinus,' 'entasis' 'front view of an ancient Greek temple, with two columns on each side and one door in the middle. Doric order' (processing by the authors).

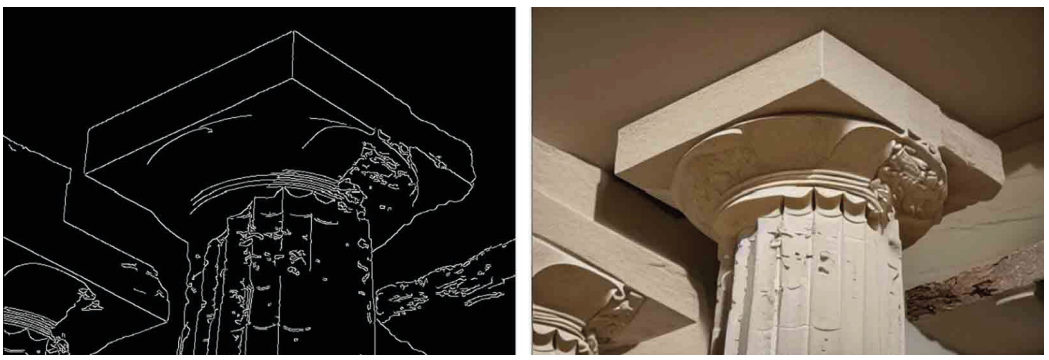

Fig. 6. Images generated with *Stable Diffusion* and *ControlNet* using the prompt 'Doric capital' and a reference image. On the left: the contour detection with *Canny edges* from *ControlNet*; on the right: the generated image (processing by the authors).

Some more accurate results have been obtained with the '/describe' function and less specific prompts, but the lack of adequate training data continues to cause bias in the output (fig. 4).
Finally, *Stable Diffusion* also has similar problems: since it does not have precise references on the Doric order, it produces approximate images. The only partially correct result was obtained by generating the facade of a Doric temple, albeit with apparent errors (fig. 5).
Integration with *ControlNet* has improved the accuracy of the generated images, consequently reducing hallucinations. In this case, the system is less affected by training data (erroneous or deficient). It relies more on replicating the contours of a pre-existing image through the *Canny edge* function. It introduces stakes in defining the shape but does not imply a correct representation of the Doric order (fig. 6).
In general, all platforms tested showed a lack of specific training data on the Doric order, both textual and visual. However, implementing such datasets poses the accessibility problem of 'commercial' Neural Network data.

Critical analysis

The study has shown how the Neural Network, built to simulate the logical processes inherent in human reasoning through inferences, can produce results inconsistent with the demands without adequate prompts and source information. The experimentation revealed the importance of defining correct text descriptions, a necessary but insufficient condition for eliminating hallucinations. This lack of connection between input and output, mediated by an external model with specific functionality, demonstrates how such tools are unreliable. This output variability is interesting in the design domain, in which a creative process receives positive stimulation from constructing unexpected inferences. However, in architectural representation, it is essential to generate formally correct content of scientific value; otherwise, there is a risk of misinterpretations of reality and the transmission of wrong knowledge.

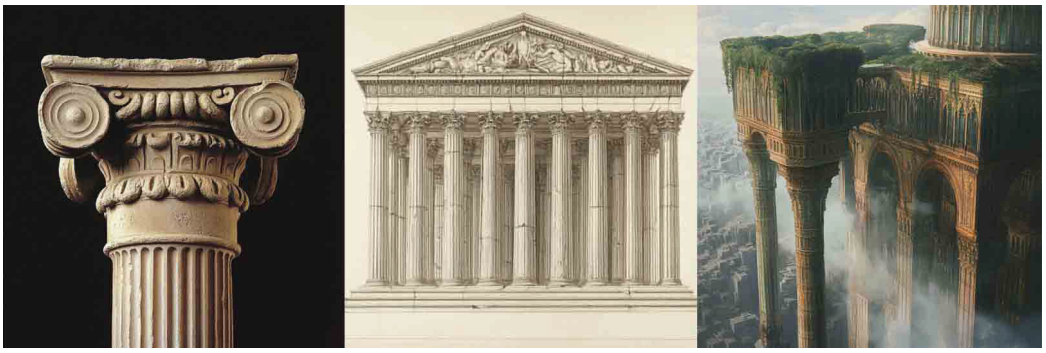

Fig. 7. Three levels of generative hallucination. From left to right: mild, intermediate, and complex. (Processing by the authors).

Paradoxically, while LLM models turn out to be very reliable, the lack of a direct relationship with LVLM models leads to a significant frequency of hallucinations. If one considers this phenomenon only as an error, metrics can be applied to quantify the percentage of correct or incorrect results against a baseline grid. However, when a baseline grid is missing, it becomes more complex to determine the error level. In the case of architectural representations, it is not easy to establish precise metrics. Three levels of hallucination can be identified generically (fig. 7):
1. mild: generated in the shallower layers of the Grid, with input data partially present. Minor errors, such as inconsistent details, minimal textural discontinuities or perspective inaccuracies manifest. These errors are visible only to careful and experienced observers;
2. intermediate: derived from less input data and greater depth of Network processing. It introduces evident errors such as the absence of architectural elements, the presence of shapes with no structural sense, and distorted perspectives. This hallucination can cause a misreading of the result. A user without specific knowledge may still perceive the content as consistent;
3. complex: generated in the deepest layers of the network with little or no input data. It produces unreal content without entire architectural portions, severe formal discontinuities, and completely inconsistent contexts, easily recognised even by inexperienced users.

Conclusions and future prospects

The research shown aimed to verify the current capabilities of some generalist text-to-image Neural Networks to produce scientifically correct images, focusing on the domain of Doric order representation. The study showed that, at present, such Neural Networks are still unable to generate scientifically reliable images, producing artefacts called 'digital hallucinations.' Using balanced and well-structured datasets and implementing specific models and continuous testing phases in which human supervision is essential to filter, correct and validate the results can lead to reducing this phenomenon. These aspects highlight the need to systematise different skills to achieve a meaningful result.
The output obtained from the experimentation made it possible to outline future research scenarios. In the short term, one can focus on improving the language and textual references by introducing those technical terms that facilitate searching for the few scientifically correct illustrations in the databases. In addition, introducing the original translations of the Latin and Greek treatises can aid the construction of inferences. However, the lack of adequately classified data remains the main limitation. Therefore, a second research phase should focus on defining strategies for targeted training of generalist Neural Networks. It can be done by implementing the datasets with correctly classified images and/or introducing filters that can introduce stakes in the search, improving the fine-tuning of the result. This procedure has the main limitation in the limited data accessibility of these networks. A final road consists of generating a new model specialised in generating consistent and scientifically reliable images and designing a training phase with correctly classified images inserted in a multilevel data-

base capable of relating to all possible queries in the architectural field. This path, although onerous in terms of computing resources and time, is currently the only one capable of building a product that can expand, through this technology, the possibilities of exploration offered to scholars in the field of architecture.

Acknowledgments

The writing of the paragraphs was organized as follows: Giulia Flenghi authored *State of the Art and A Comparison of Platforms*; Michele Russo developed *Critical Analysis*; Luca J. Senatore worked on *Digital Hallucinations and Conclusions and Perspectives*.

Reference List

Biten, A. F., Gómez, L., Karatzas, D. (2022). Let there be a clock on the beach: Reducing object hallucination in image captioning. In *IEEE/CVF Winter Conference on Applications of Computer Vision* (WACV), Waikoloa, pp. 2473-2482. https://doi.org/10.1109/WACV51458.2022.00253.

Del Campo, M., Leach, N. (Guest Eds.). (2022). *Architectural Design*, *Machine Hallucinations: Architecture and Artificial Intelligence*, Vo. 92, 3.

Gao, Y., Wang, J., Lin, Z., Sang, J. (2024). *AIGCs confuse AI too: Investigating and explaining synthetic image-induced hallucinations in large vision-language models*. arXiv preprint arXiv:2403.08542. https://arxiv.org/abs/2403.08542.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. (2014). *Generative adversarial networks. Advances in Neural Information Processing Systems*, 27.

Huang, W., Liu, H., Guo, M., Gong, N. Z. (2024). *Visual hallucinations of multi-modal large language models*. arXiv preprint arXiv:2402.14683. https://arxiv.org/abs/2402.14683.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Xu,Y., Ishii, E., Jin Bang, Y., Madotto, A., Fung, P. (2023). Survey of hallucination in natural language generation. In *ACM Computing Surveys*, 55(12), Article 248, pp. 1-38. https://doi.org/10.1145/3571730.

Maleki, N., Padmanabhan, B., Dutta, K. (2024). AI hallucinations: A misnomer worth clarifying. In *IEEE Conference on Artificial Intelligence* (CAI), Singapore, pp. 133-138. https://doi.org/10.1109/CAI59869.2024.00033.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A. Chen, M., Sutskever, I. (2021). *Zero-Shot Text-to-Image Generation*. arXiv:2102.12092 https://doi.org/10.48550/arXiv.2102.12092.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2021). *High-resolution image synthesis with latent diffusion models*. arXiv preprint arXiv:2112.10752. https://arxiv.org/abs/2112.10752.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., Das, A. (2024). *A comprehensive survey of hallucination mitigation techniques in large language models*. arXiv preprint arXiv:2401.01313. https://arxiv.org/abs/2401.01313.

Authors
Giulia Flenghi, Sapienza Università di Roma, giulia.flenghi@uniroma1.it
Michele Russo, Sapienza Università di Roma, m.russo@uniroma1.it
Luca J. Senatore, Sapienza Università di Roma, luca.senatore@uniroma1.it

To cite this chapter: Giulia Flenghi, Michele Russo, Luca J. Senatore (2025). Eidomatic hallucination of architectural orders in the age of Neural Networks. In L. Carlevaris *et al.* (Eds.). *èkphrasis. Descrizioni nello spazio della rappresentazione/èkphrasis. Descriptions in the space of representation.* Proceedings of the 46th International Conference of Representation Disciplines Teachers. Milano: FrancoAngeli, pp. 2777-2792. DOI: 10.3280/oa-1430-c900.

ISBN 9788835182412