Journal of Cultural Heritage 71 (2025) 175–183

Contents lists available at ScienceDirect

Journal of Cultural Heritage

journal homepage: www.elsevier.com/locate/culher

Original article

Identification of stone deterioration patterns with large multimodal models. Definitions and benchmarking

Daniele Corradetti [a,b,*], José Delgado Rodrigues [c]

[a] Grupo de Fisica Matematica, Instituto Superior Tecnico, Av. Rovisco Pais, Lisboa, 1049-001, Portugal
[b] Departamento de Matematica, Universidade do Algarve, Campus de Gambelas, Faro, 8005-139, Faro, Portugal
[c] Consultant in Conservation of Cultural Heritage, Rua Cidade da Beira, 76-4E, Lisbon, 1800-070, Lisbon, Portugal

ARTICLE INFO

Article history:
Received 8 June 2024
Revised 7 November 2024
Accepted 22 November 2024

Keywords:
Cultural heritage
Large multimodal models
Benchmarking
Stone deterioration patterns

ABSTRACT

The conservation of stone-based cultural heritage sites is a critical concern for preserving cultural and historical landmarks. With the advent of Large Multimodal Models, as GPT-4omni (OpenAI), Claude 3 Opus (Anthropic) and Gemini 1.5 Pro (Google), it is becoming increasingly important to define the operational capabilities of these models. In this work, we systematically evaluate the image classification capabilities of the main foundational multimodal models to recognise and categorize anomalies and deterioration patterns of stone elements that are useful in the practice of conservation and restoration of world heritage. After defining a taxonomy of the main stone deterioration patterns and anomalies, we asked the foundational models to identify a curated selection of 354 highly representative images of stone-built heritage, offering them a careful selection of labels to choose from. The result, which varies depending on the type of pattern, allowed us to identify the strengths and weaknesses of these models in the field of heritage conservation and restoration.

1. Introduction

The preservation of cultural heritage sites is a critical challenge in the fields of archaeology, historic preservation, and conservation science. Stone monuments, buildings, and artefacts are subject to a wide range of deterioration processes over time, including weathering, erosion, biological growth, salt crystallisation, and human-induced damage [1]. This study focuses on the task of image classification, aiming to identify and categorize various stone deterioration patterns essential for developing effective conservation strategies and interventions.

In recent years, advances in artificial intelligence and machine learning have opened up new possibilities for automated analysis of stone-built heritage. In particular, the development of large multimodal models (LMM) - i.e., AI systems that can process and generate multiple types of data, including text, images, and audio [2] - has the potential to revolutionise the way we study and preserve the world' s stone-built heritage.

Unlike their predecessors, LMMs are capable of processing and integrating multiple forms of data. This multimodal capability allows these models to understand complex inputs that involve more than one type of data, such as video, images and text, much like how humans use multiple senses to perceive and interpret the world around them. For example, when presented with an image of a historic monument showing signs of deterioration, an LMM can analyze the visual features of the image to identify deterioration patterns and, at the same time it can generate descriptive text explaining these patterns, possibly even suggesting causes or remedies based on textual information it has learned. This integrated approach improves the model's ability to perform tasks that require both visual perception and language understanding.

It is important to immediately note the difference between these models and those resulting from specific neural networks dedicated to identifying specific deterioration patterns as they were usually developed prior to 2023. Indeed, these neural networks were classical examples of narrow artificial intelligence [3] visual recognition systems created ad hoc for specific stone deterioration patterns through transfer learning from neural networks structured for specific segmentation and classification problems (most existing studies have focused on narrow, specialised tasks like crack detection or material classification, often using small datasets and custom-built algorithms cfr. [4–6]).

These models were trained on very specific deterioration patterns using dedicated neural networks for each modality (usually

* Corresponding author.
E-mail addresses: a55944@ualg.pt (D. Corradetti), j.delgado.rodrigues@gmail.com (J.D. Rodrigues).

https://doi.org/10.1016/j.culher.2024.11.017

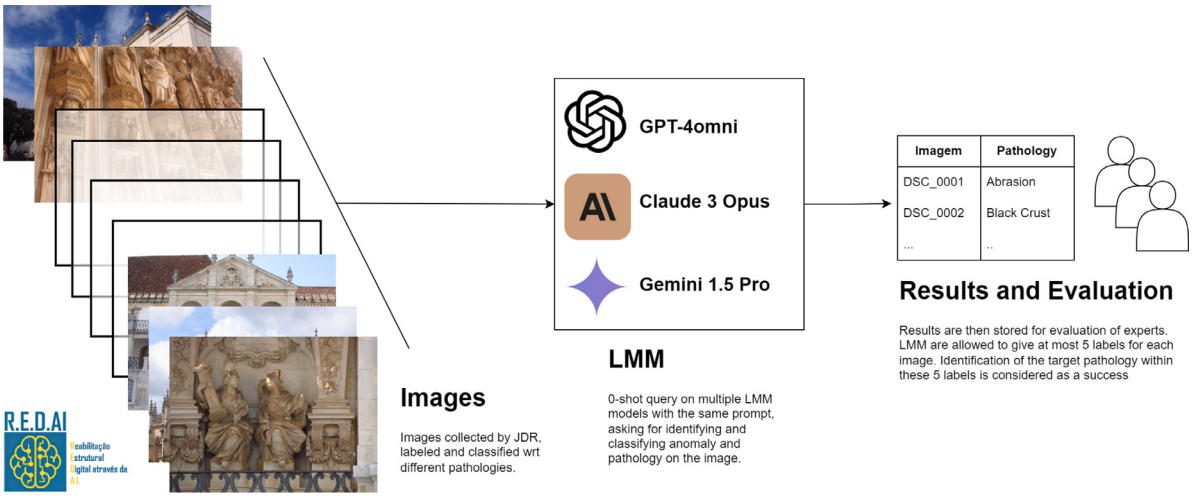

Fig. 1. Implemented workflow followed in this study.

image recognition and pattern detection) and thus lacked the ability to integrate information across different data types.

On the other hand, the current LMMs, such as OpenAI' s GPT-4omni, Anthropic' s Claude 3 Opus, and Google' s Gemini 1.5 Pro, aim to achieve a general model capable not only of recognising all specific deterioration patterns with the same model but also of performing complex diagnoses through linguistic operations or special cognitive architectures and potentially suggesting specific interventions (e.g., [7,8]).

The current generation of LMMs processes all inputs through a unified neural network that has been trained on vast amounts of multimodal data. This approach allows the model to consider the interplay between text and images naturally. For example, when analyzing a deterioration pattern, the model doesn't just recognize the visual features but also understands the terminology and descriptions associated with that pattern in literature.

Despite the extensively studied abilities of these LMMs, the specific skills and limitations of models for heritage conservation applications have not been systematically evaluated. The increasing integration of these models into human operational workflows thus necessitates an extensive and systematic evaluation of how state-of-the-art multimodal models perform on a diverse range of stone deterioration patterns and anomalies, using large, representative image sets.

2. Research aim

In this work, we aim to address this gap by conducting a rigorous evaluation of three leading multimodal models - GPT-4omni (OpenAI), Claude 3 Opus (Anthropic), and Gemini 1.5 Pro (Google) - on the task of image classification, specifically recognising and categorizing stone deterioration patterns relevant to world heritage conservation. We define a taxonomy of stone deterioration patterns based on the internationally recognised literature [9], but adapted to suit new AI interpretation tools. We then curate a dataset of 354 high-quality images exemplifying these deterioration patterns on stone heritage sites around the world. Using carefully designed prompts and label sets, we assess each model' s performance in identifying and classifying the deterioration patterns present in each image. Images and code for the test are available (see Section 5 Data Availability) in order to allow replication of the test to other foundational models or dedicated artificial cognitive entities (Fig. 1).

Finally, we want to stress out that the implemented test constitutes just a first evaluation on foundational models, without considering any finetuning or additional architectural solution around those models.

3. Material and methods

A Large Multimodal Model is an AI system capable of processing and understanding multiple modalities of information, such as text, images, audio, and video. These models aim to mimic human-like perception integrating the complementary nature of different data types. Recent advancements in deep learning, particularly transformer architectures, have enabled the development of large-scale multimodal models that can perform a wide range of tasks across various domains [10]. Multimodal models learn joint representations of different modalities [11], allowing to capture the complex relationships and dependencies between them. This enables the models to perform cross-modal reasoning, generation, and retrieval tasks [12]. Even more so, in the last generation of multimodal models, where all inputs and outputs are processed by the same neural network, in comparison with the previous generation where the multimodality was achieved by integrating different neural network each one trained on a different modality (OpenAI, 2024). The versatility of LMMs has led to their application in a wide range of domains, including healthcare, education, entertainment, and conservation science. For our purposes, applications in healthcare are inspirational, where LMMs have been used for medical image analysis, disease diagnosis, and patient monitoring [13]. While LMMs offer immense potential for automated analysis and preservation of stone monuments, buildings, and artefacts, it is also clear that before considering their application, a preliminary systematic evaluation across various deterioration categories is necessary. In our group, the opportunity to test current LMMs on such deterioration patterns was discussed. The positive response was dictated by the need to have a realistic overview of the possibility of using these tools profitably and integrating them into the concrete activities of restoration and conservation of heritage where the distinction of these patterns is still hoped for if not required. In this work, we decided to evaluate foundational multimodal models, i.e., without specific training and without additional systems or architectures, in order to recognise the native abilities of such models. The results of this study, therefore, represent a starting point and an evaluation of the native knowledge of the various models. Numerous benchmarkings were inspirational, but we mainly drew inspiration from the GPQA model [14], for which the objective of the test is to evaluate very advanced skills through the help of experts in the field. Given the multimodal nature of the

model, the choice of modality and evaluation method was taken into consideration. After careful reflection and preliminary tests, we considered that the most important and discriminating ability to evaluate would consist in visually recognising stone deterioration patterns [15]. In fact, from our preliminary analyses, we noticed that closed-form questions of a theoretical nature in the form of audio or text have a very high percentage of correct answers and do not guarantee a real applicability of the model to concrete cases. At the same time, we estimated that visual evaluation could take place optimally through the administration of photographic images representative of the various patterns to the models.

3.1. Preparation of the test

Deterioration patterns are the visible signs of the processes that continually take place in any built object. By their nature, deterioration processes are not visible and it is through the signs they impose or originate that scientists and professionals interpret and resolve them. In this sense, the correct observation and identification of existing patterns is the first and most important action that conservation professionals must take care of. Traditionally, this step is carried out through direct inspection of the object, which involves exhaustive observation of exposed surfaces and detailed documentation of all relevant data collected. Having automatic processing of observational data and correct identification of deterioration patterns would be a relevant improvement in the preparation of conservation interventions, which could end up being a high cost/benefit option that would contribute to making conservation activities more affordable and attractive. The objective of this project is to obtain an AI tool capable of correctly identifying deterioration patterns, graphically documenting their distribution throughout the object, and calculating areas and estimating costs to carry out the corresponding conservation intervention. Before embarking on designing a dedicated AI tool to specifically address the project' s objectives, we decided to systematically evaluate the performance of three of the last generation LMMs on stone deterioration patterns recognition, which required a comprehensive benchmarking study. The preparation of the test involved two key steps: creating a taxonomy of stone deterioration patterns and selecting a representative set of images for each pattern. In some preliminary trials, LMMs were allowed to use an open taxonomy, which showed that patterns were described in colloquial terms that varied from case to case, even when they described the same pattern.

3.1.1. Creating the taxonomy

To avoid ambiguity in interpretations as much as possible and to focus the description made by LMMs, a comprehensive taxonomy of stone deterioration patterns relevant to world heritage conservation has been defined. It is largely based on the ICOMOS-ISCS Glossary of deterioration patterns (ICOMOS 2008), with some adaptations to better meet the needs of this activity, following the consensus gathered in a quick consultation with some professionals in the area. After careful evaluation of the many terms currently used to describe such patterns, the following list was selected:

> *Abrasion, adherent deposit, algae, alveolisation, biological colonisation, black crust, blistering, chipping, contour spalling, corrosion of inserted elements, crack, craquele, dark diffuse biocolonisation, deformation, degraded joint filling, detachment of mortar layer, differential erosion, discolouration, efflorescence, encrustation, erosion, film, flaking, fracture, fragmentation, gap, graffiti, granular disintegration, lichens, loose deposit, misalignment elements, moist area, moss, open joint, patina, perforation, pitting, plant, powdering, soiling, spalling, staining, sugaring, thin black deposit, unaesthetic joint filling, unaesthetic patch repair.*

Fig. 2 illustrates some of the deterioration patterns used to test the models in the present study. This list covers a wide range of stone deterioration patterns, including the major categories of damage signs commonly encountered in stone-built objects: patterns linked to visual disturbances; patterns representing erosion or mass losses; patterns with direct or indirect connection with structural stability issues. As discussed later in this article, LMMs showed great difficulties in identifying some chosen patterns, but no effort was made here to adapt or change the number and type of patterns to see how such adaptations could impact the capabilities of the tested models. However, such adaptations are not to be excluded in future tests if deemed useful to benefit the models learning capacity and to improve the produced outputs.

3.1.2. Selecting the images

The next step in preparing the test was to curate a dataset of high-quality images representative of each stone deterioration pattern in our taxonomy. We started with a large archive of over 8000 images collected from various sources, including field surveys and conservation reports. The images depict stone heritage sites from around the world, spanning different historical periods, architectural styles, and geological contexts. Although they cover a wide variety of patterns, the images were not taken with this study in view and are therefore certainly not the ideal set that one could hope to test. To select the most suitable images for our testing, we established a set of criteria. First, each image had to clearly contain the "Target Pattern" as a major occurrence in its contents. Second, the images had to be of sufficient resolution and quality to allow for detailed analysis by the LMMs. Third, we aimed to include a diverse range of stone types, surface textures, and environmental conditions to assess the models' robustness and generalisation capabilities. It is worth noting that while each image was selected with a specific Target Pattern in mind, they always contain other patterns that the models were asked to identify as well. At the end of the process the total number of the selected images was 354. Each photo was selected having one main deterioration pattern sufficiently clear to be identifiable beyond reasonable doubts. Given the fact that not all patterns are equally abundant, the authors could not afford to extract from their personal archives a balanced distribution for all patterns (see Fig. 3). Since the object was not a comparison of performances across the different patterns, this unequal representation was not considered to be a problem for the extracted conclusions. Finally, it is worth noting that all images used in this study are sourced from the personal archive of José Delgado Rodrigues (JDR) and were captured during field surveys and documented conservation projects conducted by JDR over the past decade. As the sole intellectual property of JDR, these images are free from third-party copyrights, and permission has been granted to use and reproduce them for academic and research purposes as specified in the Data Availability Section.

3.2. Selection of the models

For this study, we selected three state-of-the-art foundational multimodal models released in 2024: OpenAI GPT-4omni [16], Anthropic Claude 3 Opus [17], and Google Gemini 1.5 Pro ([18]). These models were chosen based on their very strong performance on all general benchmarks (e.g., MMLU, GPQA, etc.), their ability to handle a wide range of tasks across different domains and, also, by their adherence to ethical AI practices and data protection standards. OpenAI GPT-4omni is an extension of the GPT (Generative Pre-trained Transformer) architecture, which has shown remarkable success in natural language processing tasks. GPT-4omni incorporates visual processing capabilities, enabling it to understand and generate both text and images. The model has been trained on a vast amount of web-scale data, processes multiple data types

Fig. 2. Example of the deterioration patterns used for benchmarking.

simultaneously, thus improving its ability to understand and generate diverse outputs. Having real-time responses, being scalable and an improved efficiency, this model is naturally one of the best candidates for deployment of real applications to cultural heritage conservation tasks. Claude 3 Opus is Anthropic' s next generation model offering enhanced capabilities in multimodal processing. Claude 3 Opus builds upon the successes of the original Claude model, which demonstrated strong performance on language understanding and generation tasks. Moreover, Anthropic is known for its focus on ethical and robust AI systems. As GTP4-omni, Claude 3 Opus processes various visual formats, making it potentially suitable for detailed heritage site analysis. Google Gemini 1.5 Pro is the most advanced model of Google' s Gemini series of multimodal models. It shows excellent results on a vast range of tasks, including image captioning, visual question answering, and image-text retrieval. Being released by Google, Gemini 1.5 Pro is probably the easiest model to scale and the most cost-effective, which thus makes it one of the best candidates for deploying robust solutions with possibly high-volume of inference requests. Indeed, Gemini 1.5 Pro has a context window of 1 million token, which would translate in a 40 min long video, the possible result of a drone inspection of a site [18]. One last remark on the selection of the models is on the data security and the ethical considerations that are paramount in the deployment of Large Multimodal Models (LMMs) for cultural heritage conservation. All the selected models, i.e., GPT-4omni, Claude 3 Opus, and Gemini 1.5 Pro, fulfil highest ethical practices and data protection standards. Moreover, all three companies offer an API service that allows the images to not be stored or retained by the service providers.

3.3. Implementation of the test

To evaluate the selected multimodal models on the task of stone deterioration patterns recognition, we implemented an automatised testing pipeline (Fig. 1) which included a simple preprocessing and a prompt engineering feature. The first step in the testing pipeline was to preprocess the curated dataset of images in a way compatible with the input size of the selected models. To get accurate and concise responses from the models, we used a prompt that allowed the model to provide as many patterns as necessary up to five, but only using the list of stone patterns defined in our taxonomy. We then utilized the official APIs provided by OpenAI, Anthropic, and Google to interact with their respective multimodal models. Along with the test image provided as the user, the following system prompt for all three models was:

```
prompt =‘‘'Given the pathologies in this list
{Pathologies}. What is the pathology or
pathologies of the stones in the picture
```

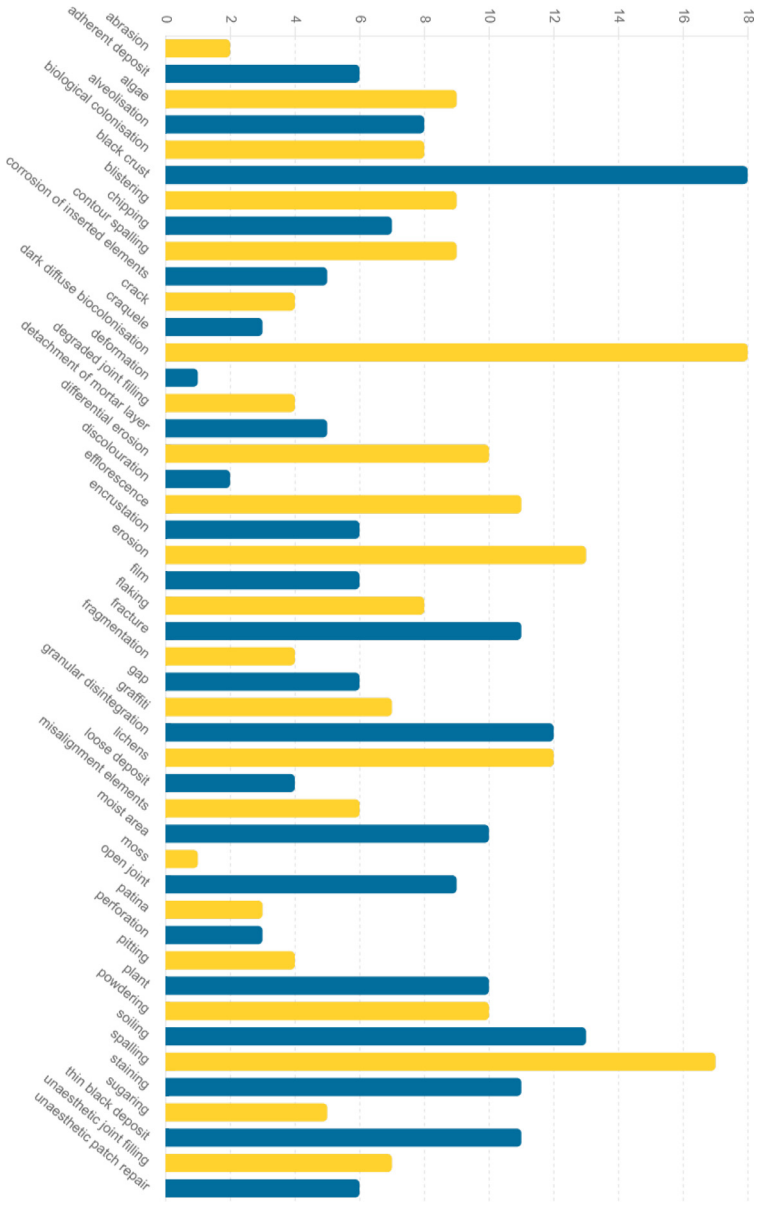

Fig. 3. Number of selected images for each deterioration pattern as follows: abrasion (2), adherent deposit (6), algae (9), alveolisation (8), biological colonisation (8), black crust (18), blistering (9), chipping (7), contour spalling (9), corrosion of inserted elements (5), crack (4), craquele (3), dark diffuse biocolonisation (18), deformation (1), degraded joint filling (4), detachment of mortar layer (5), differential erosion (10), discolouration (2), efflorescence (11), encrustation (6), erosion (13), film (6), flaking (8), fracture (11), fragmentation (4), gap (6), graffiti (7), granular disintegration (12), lichens (12), loose deposit (4), misalignment elements (6), moist area (10), moss (1), open joint (9), patina (3), perforation (3), pitting (4), plant (10), powdering (10), soiling (13), spalling (17), staining (11), sugaring (5), thin black deposit (11), unaesthetic joint filling (7), and unaesthetic patch repair (6).

```
given by the user? Answer only listing the
pathology or pathologies separated by a comma
without stating anything else.'''.
```

The model's response, containing the identified patterns, was then received and processed. Finally, the image filename, true pattern label, and the model' s identified patterns were stored in a results list for the performance evaluation.

Table 1
Success rates i.e., the recall metric here expressed in percentage, on the identification of the target deterioration pattern.

Model	*Targets correctly identified*	*Total possible*	*Success rate (%)*
GPT-4omni	**149**	354	**42.1%**
Gemini 1.5 Pro	138	354	39%
Claude 3 Opus	86	354	24.3%

4. Results

After the model response, the results were collected and then manually validated. The first action was to confirm whether or not the "target pattern" was one of the identified patterns. Then, the original image was opened to check which of the identified non-target patterns were actually present in the image. The first relevant result is expressed in the rate of success i.e., recall, the models showed in identifying the "Target Pattern" (Table 1 and Fig. 4). Indeed, it is important for us to be sure that the most evident or relevant deterioration pattern is systematically identified by the LMMs.

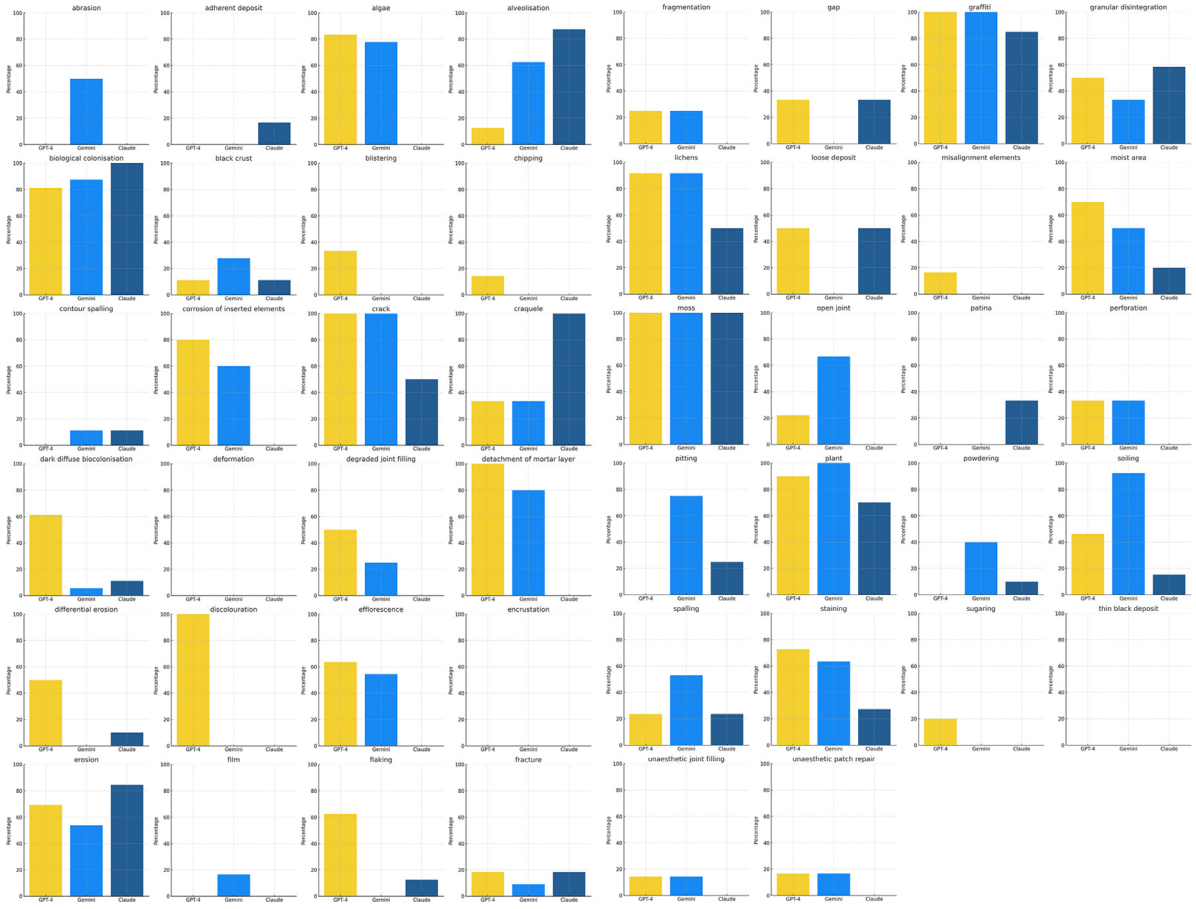

Fig. 4. Rate of success, i.e., the recall metric here expressed in percentage, of each model on the "Target deterioration pattern". *Yellow* is GPT-4omni, *light blue* Gemini 1.5 Pro and *blue* is Claude 3 Opus. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Precision of the models in the identification of non-target patterns.

Model	*Patterns correctly identified*	*Total identification*	*Success rate (%)*
GPT-4omni	677	1031	65.6%
Gemini 1.5 Pro	**741**	**1066**	**69.5%**
Claude 3 Opus	745	1263	58.9%

Besides the capability to identify the target patterns, it was also considered as relevant to know the precision in identifying non-target patterns present in the images, as openly taken by the model from the prescribed list of patterns (Table 2 and Fig. 5). Finally, in order to provide additional insights we conducted an exploratory segmentation task to assess the models' ability to localize the identified deterioration patterns within the images (Fig. 6). Although this task is not intended as a validation step for the classification results, we considered it relevant for possible future studies on the subject helping us in understanding the models' ability related to pattern localization.

4.1. Evaluation metrics

To describe the performance of the models we employed the Recall metric, which corresponds to what we will simply call the "success rate" of the model, and which is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP are the True Positives, i.e., the correct identifications of the deterioration patterns present in the images, while FN are the False Negatives, i.e., the instances where the patterns were present but not identified by the models. This metric measures the proportion of actual positives (the deterioration patterns present in the images) that were correctly identified by the models. Recall is particularly important in scenarios where missing a positive instance is costly, such as medical diagnoses or, in our case, the identification of critical deterioration patterns in heritage conservation. For the identification of non-target patterns, we used the Precision metric, calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where FP (False Positives) are the instances where the model identified a pattern that was not actually present in the image. Measuring the proportion of positive identifications, Precision is important when assessing the reliability of the model's predictions. As can be seen, our conclusions are of a qualitative nature, and introducing more complex metrics might lead readers to expect a level of quantitative accuracy that our exploratory experiment was not designed to provide.

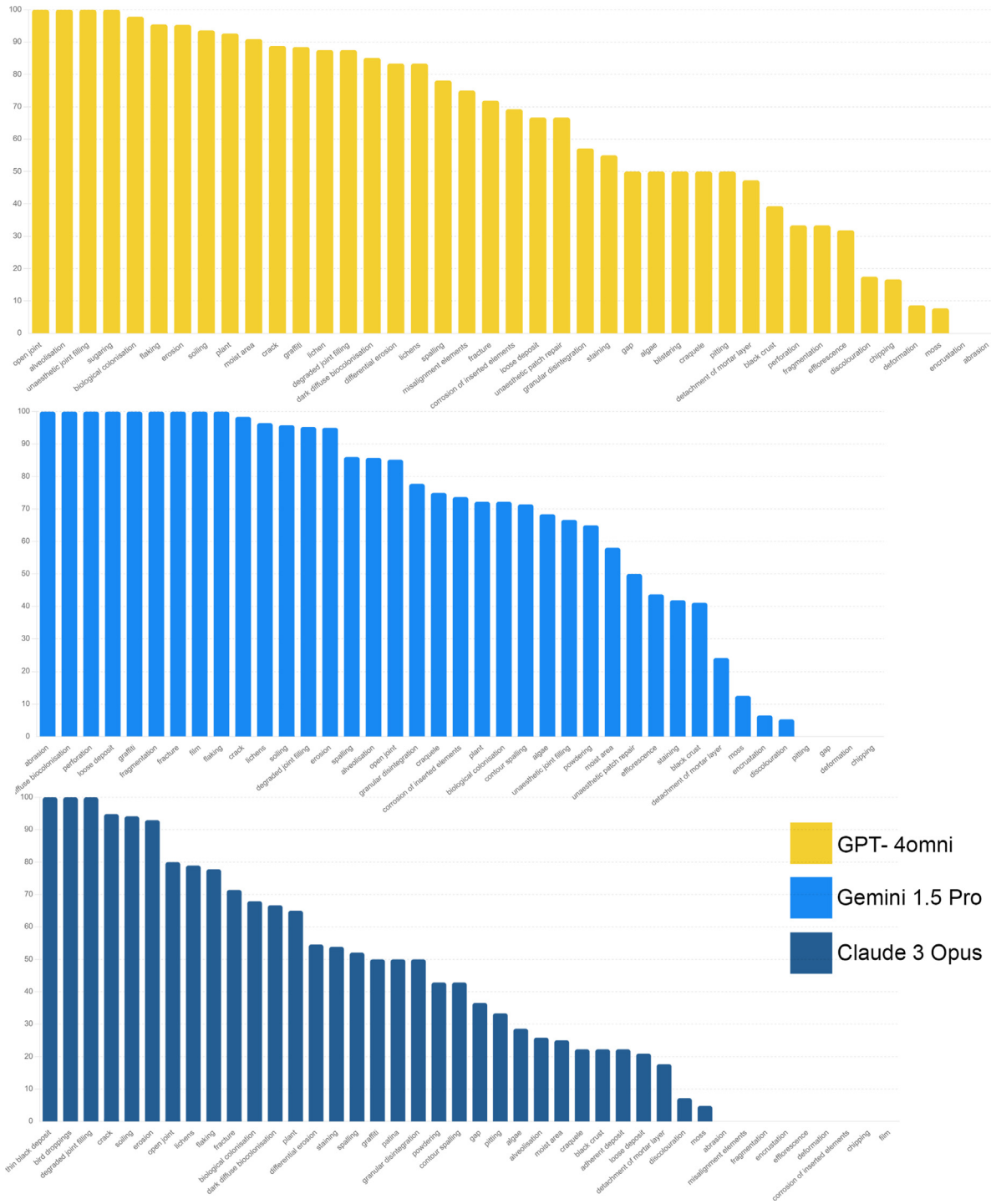

Fig. 5. Precision, expressed in percentage, in the identification of the presence of non-target patterns (as openly chosen by the models). In the image the color *yellow* represents the model GPT-4omni, the color *light blue* represents Gemini 1.5 Pro and the *blue* is for Claude 3 Opus. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Target patterns

Results on the recall metric show (Table 1) that the best performing model was that from OpenAI, i.e., the GPT-4omni model, which identified the target deterioration pattern in 42.1% of the cases, followed by Gemini 1.5 Pro, which scored a 38.9% rate of success, while Claude 3 Opus scored only 24.3%. Results also show (Fig. 4) that GPT-4omni, despite its overall best performance, was not the best performing model across all target patterns. Indeed, Gemini 1.5 Pro scored way better than GPT-4omni in identifying numerous patterns (e.g. “abrasion”, “alveolisation”, “open joint”, etc.) and even Claude 3 Opus, the least performing, out passed it in identifying “craquele” and “alveolisation”. To a certain extent, this can be taken as an indication of the lack of reliability of the models for this specific use.

As a general rule of thumb, we saw that LMMs are better in the spontaneous identification of non-target patterns than in identifying the target pattern allocated to each image (Table 2 and Fig. 5).

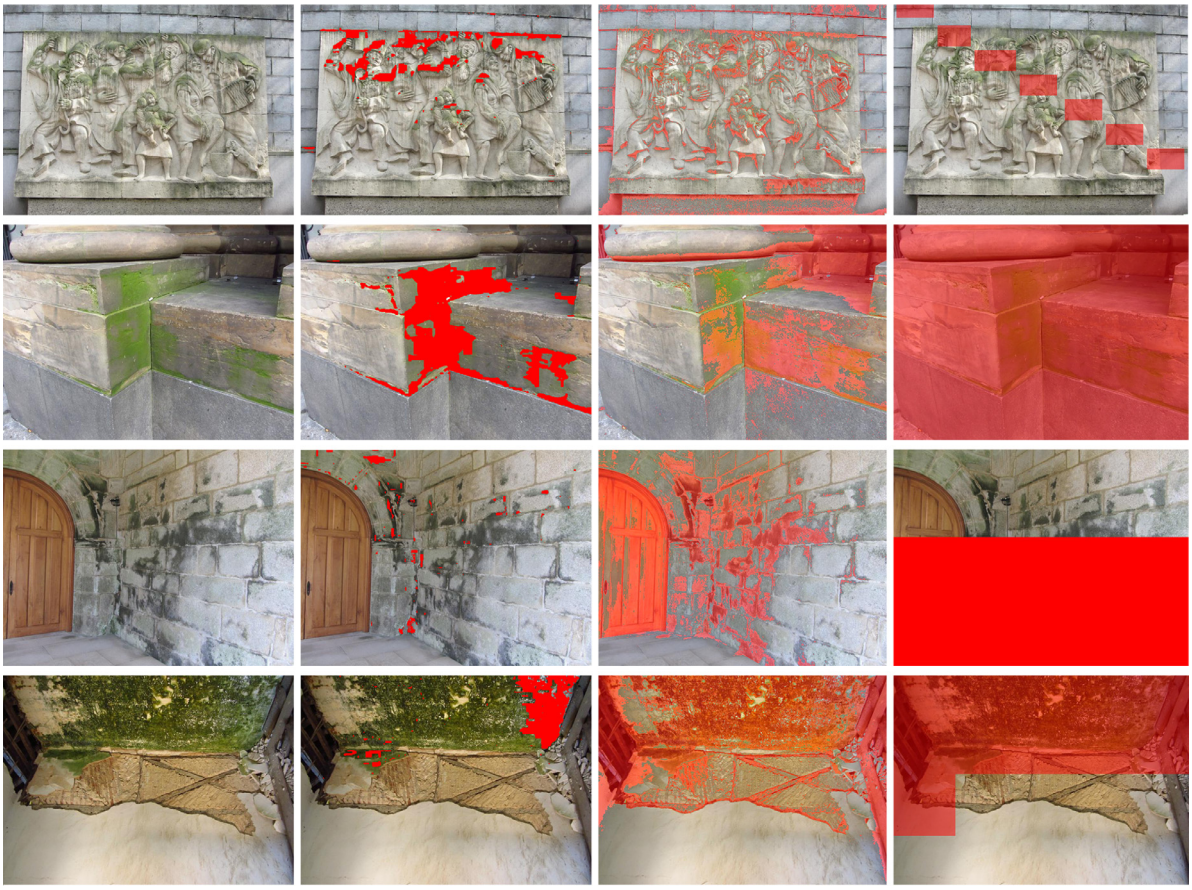

Fig. 6. Example of additional queries to visualise the answers reproducibility.

The first set of answers in Table 4 suggests that the models are still far from being reliable and that they fall short in guaranteeing that they are able to identify any specific deterioration pattern. On the other hand, the results in Table 2 show that the models already have a vast vocabulary and identification processes incorporated, although they still seem not sufficiently structured and robust, meaning that, for the most part, they are unusable for practical purposes.

Overall, these first results mean that existing models, despite their enormous capabilities, are not prepared to solve problems in the practice of conservation and restoration of built heritage. In fact, for any image identified as a paradigm of a given pattern of deterioration, only when the "success rate" (the specific metric would depend on the specific aim of the system) falls very close to 100% will it be considered ready to be used for practical purposes.

Besides the rate of success in identifying the target pattern, it is also of interest the models' overall precision in identifying the presence of patterns freely taken from the given list. In this case Gemini 1.5 Pro scored a better result than GPT-4omni. It is also worth noting that Claude 3 Opus has 9 accurately chosen patterns, but 6 of them were not in the short list of allowed patterns, for instance "window", "eroded bricks", etc. In strict terms, these hallucinations could be taken as erroneous identifications, but we opted to leave them as produced to illustrate this peculiar behaviour of the model.

4.3. Consistency and reliability

An important element in developing and deploying reliable and robust AI models for the conservation and restoration of stone involves understanding the consistency and reliability in diagnostic capabilities. This is particularly true for LMMs, whose results are not deterministic but have a certain degree of variability depending on the ' temperature' at which the model is queried. To better understand the process of pattern recognition and its robustness, we asked GPT-4omni to segment in red the areas affected by the pattern for which the model had obtained a high identification score, i.e., the presence of algae in the illustrated example (Fig. 6).

Clearly, this task serves only to provide additional insights into the models' capabilities and limitations but is not intended as a validation step for the classification results. Indeed, at the date of the test (May 2024) it was not possible to use all three models for concrete segmentation tasks or at least in a way that was homogeneous for all three models, therefore nothing more than an exploratory and qualitative test was possible.

In this figure, the first image is the original, part of the test, while the other three are the results obtained in three separate sessions with the same prompt, i.e.,

> prompt="'Colour in red the parts that are subject to algae pathology"'.

As can be seen, the model is not consistent for the same prompt, providing very different results. On one hand, this is encouraging, suggesting that there is significant room for improvement by providing the model with specific fine-tuning, precise instructions, and diagnostic guardrails. On the other hand, our analysis highlights the need for the construction of a specific cognitive architecture, in the absence of which the results of the foundational model appear to be of weak consistency and thus of questionable reproducibility.

5. Discussion and conclusions

The study carried out with a selection of deterioration patterns typically found in built heritage objects allowed us to verify that the LMMs tested here were not specifically trained for the conservation and restoration environment of built heritage. Despite the enormous capabilities that can be attributed to them, the success rates in identifying patterns are still far from what will be required of them as working instruments to produce usable results. The results obtained are extremely informative, showing a significant variation in accuracy depending on the pattern analysed. While some patterns are identified with impressive accuracy, others seem invisible to LMMs or indistinguishable from other patterns. A hypothesis formed in our work group is that the lack of specific training has led large language models (LLMs) to adopt the common, non-technical meanings of the words used to identify these patterns. The training of LMMs involves the semantic association between words and images through a process of text, video, and audio encoding into the same semantic space, then training often uses supervised learning and contrastive learning strategies [12]. However, the absence of a special attention to the technical meanings that certain words have in the context of conservation of cultural heritage leads to a confusion of terms and thus to a generic and unspecific diagnosis. A structural problem is then the difficulty of distinguishing certain specific patterns from mere photographs, for instance when a tactile evaluation is required for their identification. The three LMMs tested have a broad vocabulary relevant to the area, but lack the “understanding” of concepts and terminology specific to the conservation and restoration domain. Therefore, a specific improvement process will be necessary to bring them to a level of proficiency compatible with the needs of professionals in this field of activity. In fact, we think that the results were encouraging, suggesting that specific training sessions on dedicated datasets could improve drastically the results. Finally, it is worth noting that, given the rapid evolution of the field, the aforementioned LMMs have been trained on a large portion of the existing non-synthetic data. Given the scarcity of original data, subsequent generations of LMMs are increasingly oriented towards training-based on synthetic data generated by previous versions of LMMs [19–21]. While this type of training can be very effective in teaching general skills to the models, it is possible that a cognitive weakness in the training of the actual generation of LMMs will likely remain influencing the successive generations, if not properly addressed.

Data avalaibility

All images used in this study are sourced from the personal archive of José Delgado Rodrigues (JDR). Permission has been obtained from JDR to use and reproduce these images for research and publication purposes. The curated dataset and the accompanying code are available at the following URL for replication and further studies:

https://GitHub.com/DCorradetti/REDAI_Id_Pattern

Acknowledgements

This study was carried out under the REDAI project and the Authors thank all team REDAI for the input and fruitful discussions but most of all José Paulo Costa for the conceptualisation of the project.

References

[1] E. Doehne, C.A. Price, Stone Conservation: An Overview of Current Research, Getty Publications, 2010.
[2] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, 2023. ArXiv:2306.13549. https://doi.org/10.48550/arXiv.2306.13549.
[3] M.R. Morris, J. Sohl-dickstein, N. Fiedel, T. Warkentin, A. Dafoe, A. Faust, S. Legg, Levels of AGI: operationalizing progress on the path to AGI, 2023. ArXiv preprint arXiv:2311.02462.
[4] A. Camara, A. de Almeida, D. Caçador, J. Oliveira, Automated methods for image detection of cultural heritage: overviews and perspectives, Archaeol. Prospect. 30 (2) (2023) 153–169, doi:10.1002/arp.1883.
[5] M. Mishra, P.B. Lourenço, Artificial intelligence-assisted visual inspection for cultural heritage: state-of-the-art review, J. Cult. Heritage 66 (2024) 536–550, doi:10.1016/j.culher.2024.01.005.
[6] N. Oses, F. Dornaika, A. Moujahid, Image-based delineation and classification of built heritage masonry, Remote Sens. 6 (3) (2014) 1863–1889, doi:10.3390/rs6031863.
[7] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, D. Amodei, Language models are few-shot learners, 2020. ArXiv preprint arXiv:2005.14165.
[8] K. Saab, T. Tu, W.-H. Weng, R. Tanno, D. Stutz, E. Wulczyn, F. Zhang, T. Strother, C. Park, E. Vedadi, J.Z. Chaves, S.-Y. Hu, M. Schaekermann, A. Kamath, Y. Cheng, D.G.T. Barrett, C. Cheung, B. Mustafa, A. Palepu, V. Natarajan, Capabilities of gemini models in medicine, 2024. ArXiv. https://arxiv.org/abs/2404.18416.
[9] ICOMOS-ISCS, Illustrated glossary on stone deterioration patterns, 2008. ICOMOS International Scientific Committee for Stone (ISCS).
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[11] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, 2018. ArXiv preprint arXiv:1810.04805.
[12] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: In International Conference on Machine Learning, 2021. https://api.semanticscholar.org/CorpusID:231591445
[13] Gemini Team, Advancing multimodal medical capabilities of gemini, 2024. ArXiv preprint arXiv:2405.03162. https://arxiv.org/pdf/2405.03162.
[14] D. Rein, B.L. Hou, A.C. Stickland, J. Petty, R.Y. Pang, J. Dirani, J. Michael, S.R. Bowman, GPQA: a graduate-level google-proof q&a benchmark, 2023. ArXiv preprint arXiv:2311.12022. https://arxiv.org/abs/2311.12022.
[15] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, VQA: Visual question answering, in: Proceedings of the IEEE International Conference on Computer Visio, 2015, pp. 2425–2433, doi:10.1109/ICCV.2015.279.
[16] 2024. OpenAI (2024, May 13) Hello GPT-4o, https://openai.com/index/hello-gpt-4o/retrieved May 27.
[17] Anthropic., Introducing the next generation of claude, 2024. https://www.anthropic.com/news/claude-3-family, retrieved May 27, 2024.
[18] Google, Introducing gemini 1.5, google's next-generation AI model, 2024. https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/ retrieved May 27, 2024.
[19] S. Choenni, T. Busker, M.S. Bargh, Generating synthetic data from large language models, in: 2023 15th International Conference on Innovations in Information Technology (IIT), 2023, pp. 73–78, doi:10.1109/IIT59782.2023.10366424. Al Ain, United Arab Emirates
[20] S. Gholami, M. Omar, Does synthetic data make large language models more efficient?, 2023. ArXiv, abs/2310.07830. 10.48550/arXiv.2310.07830
[21] Z. Li, H. Zhu, Z. Lu, M. Yin, Synthetic data generation with large language models for text classification: potential and limitations. ArXiv, abs/2310.07849. (2023)