Towards AI-Assisted Fossil Identification

Sascha Ingemey[1†], Megan Spielberg[1†], Vince Colson[1†], Isaak Eijkelboom[2,3], Georgiana Manolache[1], and Gerard Schouten[1]
[1]Fontys University of Applied Sciences,[2]University of Utrecht,[3]Naturalis
[†] Equal contribution
{g.manolache,g.schouten}@fontys.nl

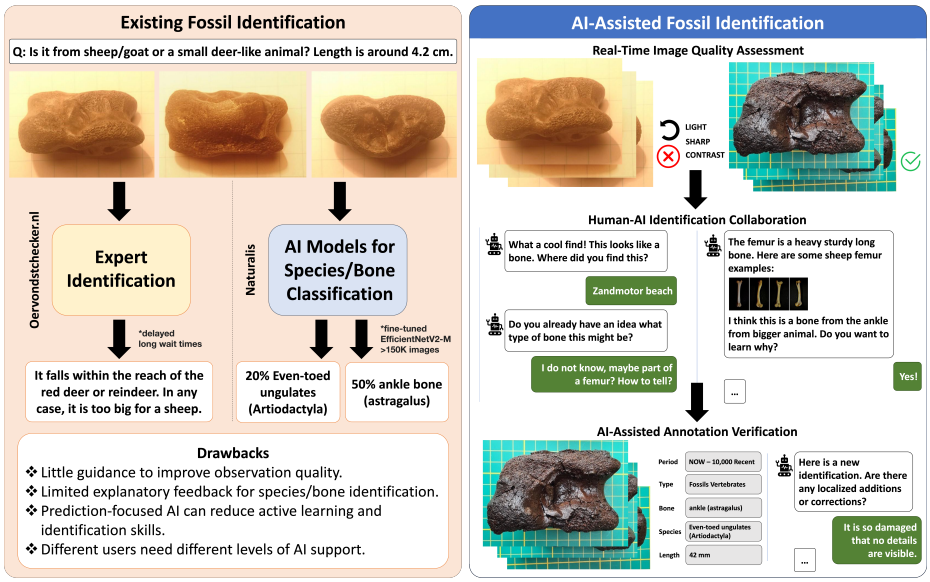


Fig. 1: **AI-assisted fossil identification overview. Left:** Existing approaches based on expert assessment or image classification. **Right:** A new paradigm combining AI-assisted guidance across identification expertise.

Abstract. Recent advances in Vision-Language Models (VLMs) offer new opportunities for paleontology, but fossil identification remains constrained by variable image and annotation quality and by subtle or fragmented diagnostic features. We investigate three directions for improving AI-assisted fossil identification: (1) real-time image-quality guidance, (2) conversational AI for citizen-science identification and data collection, and (3) richer localized morphological annotations. We introduce GoldenFossils, a proof-of-concept dataset of expert-validated localized annotations created through an AI-assisted human-in-the-loop framework, and evaluate state-of-the-art VLMs with and without localized evidence. Results show benefits from image-quality guidance, complementary strengths of conversational and form interfaces, and model-dependent gains from localized expert evidence. Together, these findings highlight how better fossil data collection can support more reliable AI-assisted identification and research into past biodiversity and evolution.

1 Introduction

Artificial intelligence (AI) is increasingly being explored for scientific identification tasks. In paleontology, recent work shows growing interest in machine-learning approaches, but also emphasizes that progress is constrained by the scarcity of large, high-quality, well-labeled datasets [43]. Similar challenges occur in machine-learning-assisted citizen-science projects, where useful AI workflows depend not only on model development, but also on the careful collection, annotation, validation, and reuse of data contributed by experts and non-expert volunteers [10]. Fossil identification is particularly challenging because taphonomic processes often leave specimens fragmented, abraded, mineralized, or stained, obscuring diagnostic morphological features. Moreover, morphological convergence can make skeletal elements from different taxa visually similar, especially after degradation. Figure 2 illustrates this challenge, showcasing five specimens that exhibit near-identical visual characteristics but belong to entirely different taxa or anatomical regions. For non-experts, such specimens can therefore be difficult to distinguish reliably. As growing numbers of fossils are collected through citizen-science initiatives, such as FossilFinder [36] or Oervondstchecker.nl [28], there is an increasing need for AI-assisted approaches that support reliable fossil image annotation and identification without requiring specialized expertise.

Visual Question Answering (VQA) offers a user-friendly approach to image-based queries, and state-of-the-art Vision-Language Models (VLMs) show promising capabilities across scientific domains [6,7,22,39,45]. However, their application to fossils remains limited. Current models (i.e., EfficientNetV2-M) achieve only 50% accuracy despite training on >150K fossil images from community-science and standardized collections, with substantially better performance on standardized museum data [9] (see Figure 3). This gap reflects both variable image quality and limited identification context. Standardized collection and citizen-science datasets typically provide taxonomic or anatomical labels but lack fine-grained morphological annotations (see Table 6). Richer descriptors (e.g., texture, color, size) are optional or inconsistent in community-science platforms such as Oervondstchecker.nl [28] and Fossilfinder [36]. Without context, VLMs may produce spurious predictions driven by biases and patterns in their training data rather than the available visual evidence [13] (see Appendix A).

Fig. 2: **Examples of fossils from GoldenFossils from different animals | anatomical parts. Left-to-right:** (1) *Cervus elaphus* (red deer) | astragalus, (2) *Cervus elaphus* (red deer) | astragalus, (3) *Bovidae* (bison) | scaphoid, (4) *Equus ferus* (wild horse) | molar, (5) *Cervidae elaphus* (red deer) | phalanx. Source: Oervondstchecker.nl [28].



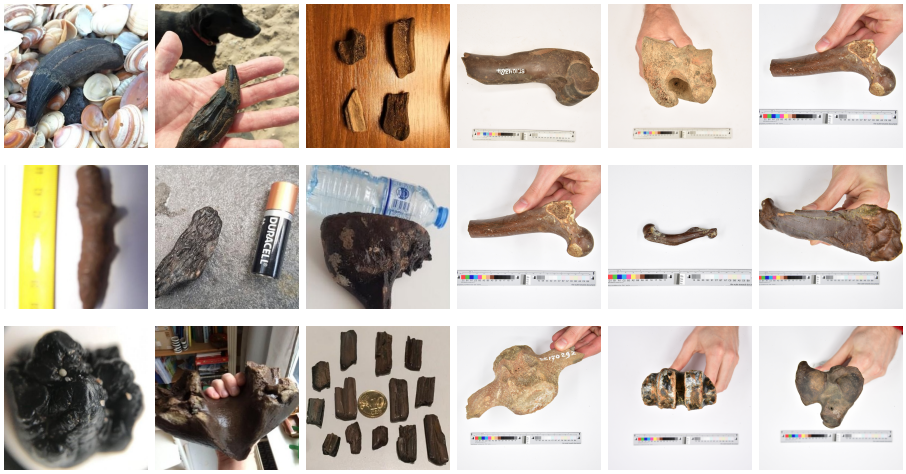


Fig. 3: **Examples of fossil images from citizen scientists and standardized museum archive. Left-to-right:** (column 1-3) submissions from Oervondstchecker.nl [28] collected from the coastline of the Netherlands; (column 4-6) archived fossils from Naturalis Biodiversity Center [23].

In this work, we explore AI-assisted fossil identification beyond image recognition, aiming to support both expert paleontologists and citizen scientists (see Figure 1). Our contribution is three-fold: (1) we investigate real-time AI guidance for improving fossil image acquisition quality; (2) we study conversational and structured AI interfaces for supporting citizen scientists in fossil identification and data collection; and (3) we introduce GoldenFossils, a proof-of-concept dataset of expert-validated localized morphological annotations created through an AI-assisted human-in-the-loop framework, and evaluate state-of-the-art VLMs with and without this localized evidence. Our results show that image-quality guidance can improve specific aspects of fossil photography, conversational and structured interfaces provide complementary benefits, and localized expert evidence can improve VLM identification, although gains are model-dependent. Together, these findings highlight the potential of moving AI beyond a standalone prediction engine toward an adaptive, dialogue-based collaborator that supports improved fossil data collection, more reliable identification, and a deeper understanding of past biodiversity and evolution.

2 Methodology

2.1 Image quality assessment

Prior work identifies fossil image quality and variability as important factors affecting machine-learning identification performance over citizen-science data [9,10]. Figure 3 illustrates the difference between a typical citizen science fossil submission and a professionally archived museum specimen.

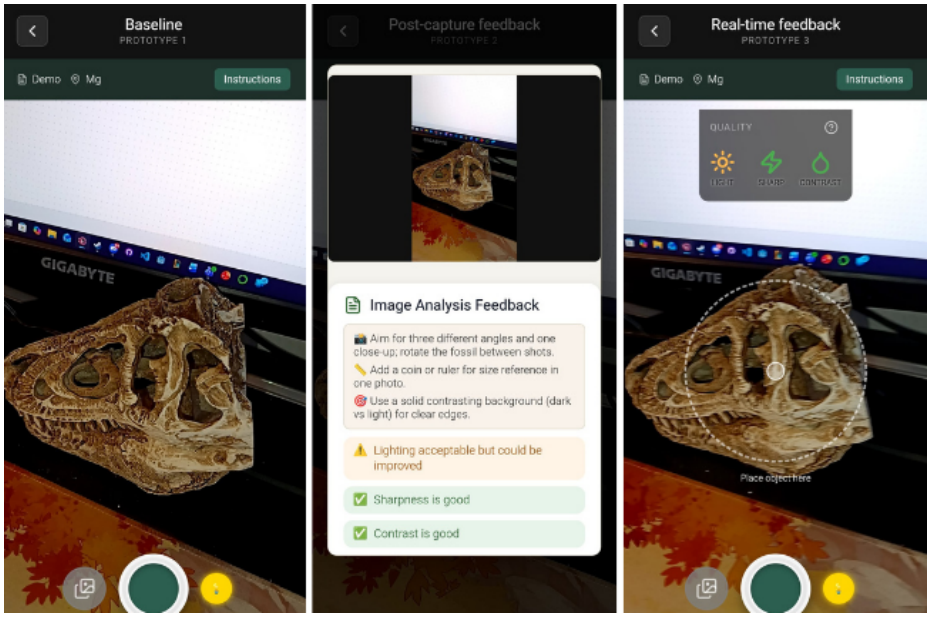


Fig. 4: **Prototype interface design for collecting high quality fossil images. Left:** Baseline prototype. **Center:** Post-Capture prototype. **Right:** Real-Time prototype.

We developed a reproducible quality metric calibrated on 100 randomly sampled Oervondstchecker.nl images [28], rated from 1 (very poor) to 5 (very good) by an expert paleontologist. Three reference-free metrics—**lighting** (mean grayscale intensity) [16], **sharpness** (Laplacian variance), and fossil-background **contrast**—were calibrated against expert ratings to define five quality levels. Scale and viewing-angle quality were assessed manually (see Appendix D).

A comparative experimental design was used to measure and evaluate the effects of various feedback mechanisms on 183 citizen-science images from 28 participants (including paleontologists and citizen scientists). This experiment was based on three prototypes, each representing a different level of user guidance. This allows for a direct comparison between one control condition and two interventions (see Figure 4. **Baseline** prototype serves as the baseline and is the control condition based on existing platforms used for citizen science submissions (Oervondstchecker.nl [28]). In this prototype, users take and upload images without receiving any feedback on their image quality. **Post-Capture** prototype provides feedback on image quality after an image has been taken and encourages users to retake images based on that feedback. **Real-Time** prototype offers immediate feedback during the image-capturing process, guiding users to take optimal images on their first attempt.

For each prototype, testers followed the same process: (1) taking one or more images; (2) reviewing the provided feedback mechanism; (3) being able to delete any unwanted images before submission; (4) submitting their final selection of

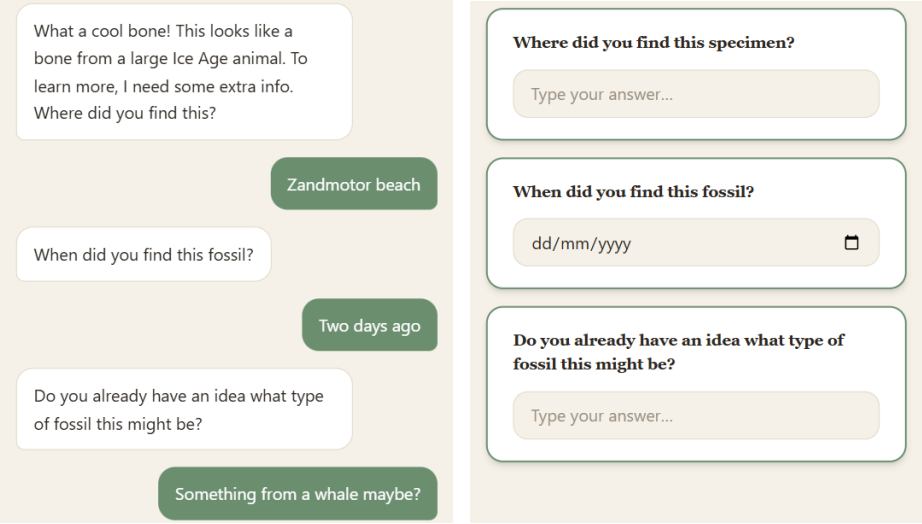


Fig. 5: **Conversational vs. form-based AI-assisted interfaces for fossil identification.** Participants were randomly assigned to an interface order and fossil case, completed both interface conditions, and then filled in a final comparative survey. Survey is available via ai-fossil-assistant.vercel.app.

images. Each tester evaluated all three prototypes, and the order in which they were tested was determined by balanced randomization to mitigate any order bias.

2.2 Conversational vs. form-based AI-assisted interfaces

Recent research on AI-assisted conversational interfaces suggests they lower participation barriers and improve engagement compared with traditional forms, but often at the cost of efficiency, structure, and predictability [18,31,35]. However, they may also introduce ambiguity, context-management challenges, and risks of over-reliance [11,15]. We therefore conducted a within-subject study with 51 participants, from novices to experts, comparing **conversational interfaces** using Google Gemini 2.5 Flash-Lite [12] with prompts and quick replies and **structured step-by-step forms** (see Figure 5). Participants completed both conditions and evaluated them through post-task and comparative questionnaires (see Appendix E).

Two scripted fossil-identification cases were used in the study flow. Each case was represented by three standardized images showing the fossil from different views. The first case represented a large, flat fossil bone that led to the scripted identification of a *Mammuthus primigenius* (woolly mammoth) scapula. The second case represented a long, robust fossil bone that led to the scripted identification of a *Equus caballus* (wild horse) metatarsal. The two cases were selected to represent different fossil-identification challenges. The mammoth case required participants to inspect broad shape, perceived weight, surface structure,

Fig. 6: **Examples of localized annotations from GoldenFossils collected with the AI-assisted annotation framework.** (left) molar of a *Equus ferus* (wild horse) annoated 1-typical linguaflexid of horses, 2-ectoflexid does not penetrate isthmus; (middle) lamelle of a molar of a *Mammuthus primigenius* (woolly mammoth) annoated 1-overall thin and wide; (right) jugal bone of a *Mammuthus primigenius* (woolly mammoth) annoated 1-very large trabeculae to keep the skull light.



and joint-related morphology. The wild horse case required attention to long-bone proportions, thickened ends, surface smoothness, color, and fossilization. In both cases, the questions were designed around observable characteristics that non-expert participants could describe, rather than expert anatomical terminology. The scripted questions were reviewed by a fossil expert to ensure domain relevance while preserving accessibility for non-expert participants.

Because all participants experienced both interfaces and both fossil cases, the presentation order of each case was randomly varied to support within-participant comparison while partially limiting order-related bias [5]. At the end, participants completed a comparative survey consisting of a forced-choice preference question between the two interface types, an open-text explanation of that preference, and an optional email field for follow-up communication regarding the study results.

2.3 GoldenFossils

Multimodal large language models (MLLMs) such as GPT-4o [29] have demonstrated a powerful ability in tasks involving images and natural language like captioning, VQA, and open-ended reasoning. We explore MLLMs, evaluating their ability to identify fossils and generate descriptive captions that capture relevant morphological and taxonomic characteristics. Building on existing AI-assisted annotation workflows, we use a VLM—Google Gemini 2.5 Flash-Lite [12]—as a generative assistant to reduce the effort required to produce initial annotations while retaining expert validation and refinement [3,8,41]. A significant focus was placed on the ability to create localized annotations directly on the image (see Figure 6) [40]. Localized descriptions may help preserve the spatial link between a diagnostic feature and the specific image region in which it appears. This can make annotations more interpretable for experts and learners,

Table 1: **Descriptive statistics for image quality metrics across three feedback mechanisms over n=183 images.** Baseline (n=55) with no image-quality feedback; Post-Capture (n=63) provides feedback after capture and encourages retakes; Real-Time (n=65) provides real-time guidance during image capture. Contrast differs significantly across prototypes (Kruskal-Wallis H-test, $H = 9.03$, $p < 0.05$), with a significant difference between Baseline and Real-Time (Mann-Whitney U tests with Bonferroni correction, $U = 1236.00$, $p < 0.025$).

	Baseline		**Post-Capture**		**Real-Time**	
Metric	**Mean (SD)**	**Median**	**Mean (SD)**	**Median**	**Mean (SD)**	**Median**
Lighting	141.84 (29.83)	145.84	143.71 (26.66)	145.18	145.34 (24.49)	145.17
Sharpness	264.29 (349.06)	147.27	240.56 (233.37)	183.34	225.73 (254.57)	159.58
Contrast	27.48 (21.76)	18.63	33.42 (23.98)	31.85	38.43 (20.84)	41.88
Scale rating	4.27 (1.41)	5.00	4.19 (1.41)	5.00	4.38 (1.32)	5.00
Angle rating	4.18 (1.07)	5.00	4.21 (0.97)	5.00	4.23 (0.91)	4.00

support finer-grained comparison between specimens, and provide AI models with explicit visual evidence rather than only global image-level labels. Such annotations may therefore improve both fossil identification and the transparency of model predictions. This feature addresses a known limitation in current museum documentation software, where taxonomic signals are often lost in free-text fields [38] (see Appendix F).

Data collection. GoldenFossils, the proof-of-concept dataset generated for this work, consists of a taxonomically diverse array of N=20 fossil fragments, primarily focused on fossils from the coastline of the Netherlands (see Table 9 and Figure 10). Ten domain experts were specifically tasked with annotating the specimen images from a dataset sampled from Oervondstchecker.nl [28]. These entries consist of an image and small note to the future reviewer, typically lacking morphological context. The structure of the resulting dataset can be seen in Table 9 and Figure 17. Combined, these components provide a machine-actionable repository of expert justifications that can be used for future model fine-tuning (see Appendix B).

Benchmark. Recent studies have shown that while VLMs demonstrate emergent capabilities in general VQA tasks, they often struggle with the global vs. local visual reasoning required in specialized scientific domains [19]. The baseline capabilities of VLMs are tested by presenting them with the same data and questions that the experts were presented with. A systematic comparison is conducted between the autonomous AI output and the GoldenFossils ground truth created by the experts. This research measures the expert gap—the specific morphological nuances (i.e., subtle anatomical articulations) where human intervention was required. This analysis provides the empirical evidence needed to determine the efficiency and reliability of human-AI collaboration in paleontological labeling and documentation. To isolate how different levels of human assistance affect AI accuracy, the benchmarking is conducted across two distinct tiers: (1) **Test A: Zero-shot (Baseline)** where models are provided with

only the raw specimen image and the original discoverer's note; prompted to identify the fossil and fill the established documentation schema; (2) **Test B: Localized feature analysis (Crop-Assisted)** where models are provided with the specific coordinates and image crops of the areas of special interest defined by experts during data collection, evaluating whether human-led visual attention significantly improves taxonomic accuracy, a method previously explored in medical imaging [20,30]. We evaluate against a representative sample of leading VLMs: Gemini 2.5 Flash-lite [12], GPT-4o [29], Qwen 3.6-Plus [32] and Claude 3.5 Sonnet [2].

3 Results

We present results from the (1) **image-quality assessment**, the (2) **AI-assisted interfaces**, and the (3) **VLM benchmark on GoldenFossils**, while additional results are presented in the Appendix D, Appendix E, and Appendix H respectively.

Real-time feedback significantly improves image contrast. Image quality was assessed using five metrics: lighting, sharpness, contrast, scale rating, and angle rating. Descriptive statistics for each metric are presented in Table 1. The results show that real-time feedback produced a statistically significant improvement for contrast quality, while no significant differences were found for lighting, sharpness, scale presence, or photographed angles. The observed improvement in contrast quality suggests that real-time guidance is effective for image characteristics that depend on immediate spatial adjustments, such as repositioning the specimen or altering the camera angle relative to the background. In contrast to lighting and sharpness, which are often limited by environmental conditions or the device settings, contrast can be directly improved through user actions when prompted during the capture. Post-capture feedback did not result in statistically significant improvements across any of the evaluated metrics. This outcome suggests that delayed feedback may be less effective in guiding users to take better images. When feedback is provided only after an image has been taken, users must interrupt their workflow, reinterpret quality metrics, and decide whether corrective action is needed. Prior research in human–computer interaction shows that delayed feedback increases cognitive load and reduces task efficiency, particularly for less technologically skilled users [27,26]. The absence of significant effects for lighting and sharpness across feedback conditions could demonstrate boundary conditions of real-time guidance. Modern smartphone cameras already perform automatic exposure and focus adjustments. This is potentially limiting the small gains achievable through user instruction alone [14]. Additionally, environmental factors such as ambient lighting and physical stability are not always under the user's control in field-based settings.

The form interfaces are faster, while the chat interfaces are natural and flexible. The study included participants across a broad range of

Table 2: **Conversational vs. form-based interfaces participant characteristics, including age-group distribution, mean tech savviness, and mean fossil expertise.**

Characteristic	Category or statistic	Value
Age cluster	Young (< 18)	10 (19.6%)
Age cluster	Young adult (18–34)	18 (35.3%)
Age cluster	Middle-aged (35–54)	12 (23.5%)
Age cluster	Senior (55+)	11 (21.6%)
Tech savviness	Mean (SD), range	5.33 (1.29), 1–7
Fossil expertise	Mean (SD), range	4.14 (1.39), 1–7

Table 3: **Conversational vs. form-based interfaces order-effects results, including sample sizes, mean differences by order group, and inferential test outcomes.**

Order group	n	Chat overall	Form overall	$\Delta_{\text{chat}-\text{form}}$	Chat time (s)	Form time (s)	Chat pref. / Form pref.
Chat first	27	5.38	5.62	-0.24	161.00	105.33	8 / 19
Form first	24	5.85	5.36	0.49	102.93	115.08	19 / 5

Note. Mann–Whitney test for overall rating difference by order: $U = 141.5$, $p = 0.0005$. Mann–Whitney test for time difference by order: $U = 496.0$, $p = 0.0012$. Fisher's exact test for stated preference by order: $p = 0.0006$, odds ratio = 0.111.

age groups, with varying levels of technological familiarity and fossil expertise (see Table 2). This diversity allowed us to explore how participants with different backgrounds experienced conversational and form-based interfaces for fossil identification and data collection, providing insights into their relative usability and suitability for different user groups. Both interfaces were evaluated positively, with no clear overall preference for either interface. The form interface was significantly faster to complete, but speed did not fully determine preference: a substantial proportion of participants preferred the slower interface. The qualitative results further clarified this trade-off, showing that the form interface was mainly valued for structure, clarity, overview, and speed, whereas the chat interface was mainly valued for natural interaction, flexibility, guidance, and engagement. Most importantly, the stratified order analysis showed that the apparent direction of the interface advantage reversed depending on which interface participants encountered first (Table 3). The results therefore suggest that the two interfaces support different strengths, and that the comparison between them should be interpreted as order-sensitive rather than as evidence for a stable overall winner.

Zero-shot reveals strong baseline performance but persistent morphological and megafauna biases. Table 4 shows that localized feature information generally improves model performance, but not consistently across models. GPT-4o benefits most from crop assistance, with AP@50 increasing from 0.568 to 0.950 and Recall@10 from 0.333 to 0.530. Gemini 2.5 Flash-Lite also improves substantially and achieves the best overall performance (combined AP@50

Table 4: **Performance on different proprietary models over GoldenFossils grouped by raw specimen image (Test A), localized feature analysis (crop-assisted) (Test B) and their combined mean.** Average Precision at k=50 (AP@50) measures the precision of ranked predictions among the top 50, while Recall at k=10 (Recall@10) measures whether relevant identifications are retrieved within the top 10 (see Appendix G); higher values indicate better performance.

	Test A		Test B		Combined	
Model	**AP@50**	**Recall@10**	**AP@50**	**Recall@10**	**AP@50**	**Recall@10**
GPT-4o	0.568	0.333	**0.950**	0.530	0.759	0.432
Claude Sonnet 4.6	0.481	0.350	0.652	0.428	0.567	0.389
Gemini 2.5 Flash-Lite	0.716	**0.500**	0.915	**0.600**	**0.816**	**0.550**
Qwen3.6-Plus	**0.807**	0.400	0.577	0.266	0.692	0.333

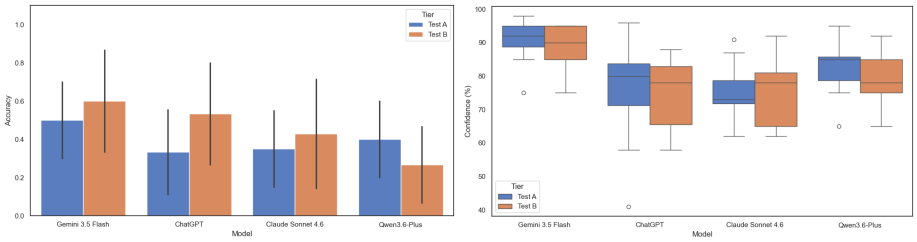


Fig. 7: **Top-1 accuracy and confidence scores over GoldenFossils grouped by raw specimen image (Test A) and localized feature analysis (crop-assisted) (Test B). Left:** Taxonomic accuracy measures agreement between model predictions and expert-verified identifications (1.0 = perfect match), with error bars indicating 95% confidence intervals. **Right:** Model-reported confidence represents the models' self-assessed certainty (%) in their taxonomic predictions.

= 0.816; Recall@10 = 0.550). Claude Sonnet 4.6 shows smaller improvements, whereas Qwen3.6-Plus performs worse with crop assistance (AP@50: 0.807 to 0.577; Recall@10: 0.400 to 0.266). Overall, the results suggest that localized morphological evidence can improve fossil identification, but its effectiveness is model-dependent. Figure 7 (left) shows the mean taxonomic match rate relative to expert-verified ground truth. In Test A (Zero-Shot), mean accuracy ranged from 0.3 to 0.5, with wide 95% confidence intervals indicating substantial variability across specimens. In Test B (Crop-Assisted), accuracy increased across all models to a mean of 0.45, while narrower confidence intervals indicated reduced variability. Figure 7 (right) shows model-reported confidence. In Test A, median confidence ranged from 75% to 90%, including high-confidence incorrect predictions. In Test B, confidence shifted slightly upward and became less variable, with scores clustering more strongly at the higher end. Generally, crop assistance can improve taxonomic accuracy without consistently increasing confidence, suggesting that confidence alone is not a reliable indicator of identifica-

Table 5: **Comparative taxonomic identification results for the baseline tier across four VLMs, illustrating autonomous baseline performance when provided only with raw specimen images and discoverer notes.** Correct identifications are highlighted in **bold**. See corresponsing images in Figure 10.

Key	Gemini 3.5 Flash	GPT-4o	Claude 3.5 Sonnet	Qwen 3.6-Plus
(a)	*Bison priscus*	***Cervus elaphus***	*Bos primigenius*	*Capreolus capr.*
(b)	***Cervus elaphus***	*Capreolus capr.*	*Capreolus capr.*	***Cervus elaphus***
(c)	***M. primigenius***	***M. primigenius***	***M. primigenius***	***M. primigenius***
(d)	***M. primigenius***	***M. primigenius***	***M. primigenius***	***M. primigenius***
(e)	*Equus ferus*	*Phoca vitulina*	*Bos primigenius*	*Bison priscus*
(f)	*M. primigenius*	*M. primigenius*	*M. primigenius*	*Callyspongia sp.*
(g)	***M. primigenius***	*Equus ferus*	***M. primigenius***	***M. primigenius***
(h)	***M. primigenius***	***M. primigenius***	***M. primigenius***	***M. primigenius***
(i)	*Plantae (Wood)*	*Indet. (Wood)*	*Cervus elaphus*	*Quercus (Wood)*
(j)	*M. primigenius*	*Bos primigenius*	***Bos/Equus***	*Pinus (Wood)*
(k)	***M. primigenius***	*Elasmobranchii*	***M. primigenius***	*Mammalia indet.*
(l)	*Mammalia indet.*	*Bos taurus*	*Mammalia indet.*	*Mammalia indet.*
(m)	***M. primigenius***	***M. primigenius***	***M. primigenius***	***M. primigenius***
(n)	***Cervus elaphus***	*Capreolus capr.*	*Capreolus capr.*	***Cervus elaphus***
(o)	*Aves indet.*	*Aves indet.*	*Mammalia indet.*	*Mammalia indet.*
(p)	*Bison priscus*	*Bos primigenius*	*Bos/Equus*	*Bison priscus*
(q)	*Bison priscus*	*Canis lupus*	*M. primigenius*	*Ursus spelaeus*
(r)	*Mammalia indet.*	*Cervus elaphus*	*Mammalia indet.*	*Mammalia indet.*
(s)	***Equus ferus***	*M. primigenius*	*M. primigenius*	*M. primigenius*
(t)	***Equus ferus***	***Equus ferus***	***Equus ferus***	***Equus ferus***

tion correctness. The results show that current models can recognize distinctive fossil morphologies but struggle with fine-grained taxonomic distinctions. Performance is strongest when specimens contain clear, characteristic features, while ambiguous or closely related taxa remain difficult. The models also show systematic biases, such as favoring well-known Pleistocene megafauna when visual evidence is uncertain. As shown in Table 5, performance was strongest for visually distinctive specimens, particularly characteristic mammoth and horse dental material, but decreased for ambiguous fragments and post-cranial elements. Collection classification was comparatively reliable, while physical size estimation remained inconsistent, especially for larger specimens without clear scale information. Overall, the results indicate that current models can recognize salient fossil morphologies but still struggle with fine-grained osteological distinctions, ambiguous material, and metric reasoning.

3.1 Limitations

While the current framework provides a technically sound environment for morphological justification, the volume of the produced dataset remains insufficient for high-fidelity VLM fine-tuning. Furthermore, the framework prioritizes the final GoldenFossils record, which can obscure the specific points where the AI failed. Future iterations of the system architecture should implement a versioning strategy within the database which could be used to audit trail whether the AI is gradually "learning" from specific categories of fossil fragments over time (see Appendix G).

3.2 Future work

Future work will focus on extending the dialogic AI framework for fossil identification, where observations are often incomplete, ambiguous, and strongly dependent on contextual evidence. A key direction is to develop systems that can dynamically identify missing evidence, request additional views, scale information, or contextual metadata, and adapt their questions to the specimen and the user's level of expertise. The framework should also support explicit comparison of alternative taxonomic hypotheses and communicate uncertainty when the available evidence is insufficient. Another important direction is the integration of expert knowledge into the dialogue process. Rather than producing a single prediction, the system could combine image-based recognition with anatomical knowledge, localized morphological annotations, and reference specimens to explain which features support or contradict a candidate identification. Future evaluations should investigate whether such interactive guidance improves both identification accuracy and the quality of citizen-science fossil observations, while also determining when expert verification should be triggered.

4 Conclusion

This work investigated AI-assisted fossil identification as a broader data-collection and human–AI collaboration challenge. Across three complementary directions, we explored how AI can support higher-quality image acquisition, accessible citizen-science interaction, and richer expert annotation. Our results show that real-time image guidance can improve specific aspects of fossil image quality, conversational and structured interfaces offer complementary interaction benefits, and localized expert annotations can improve VLM identification performance, although gains remain model-dependent. Together, these findings suggest that more reliable AI-assisted fossil identification depends not only on scaling models, but also on improving the quality and context of the data provided to them. By connecting citizen data collection, expert knowledge, and multimodal AI, this work provides a foundation for future systems in which AI helps produce better fossil data, and better fossil data in turn enables more reliable AI assistance. Ultimately, such human–AI workflows can support paleontological research into past biodiversity, environments, and evolutionary processes.

Acknowledgments

We thank domain experts who generously contributed their time and expertise to the usability and annotation study, and all the anonymous participants who took part in our qualitative evaluations (i.e., image quality Appendix D and conversational/form-based interface Appendix E).

Ethics statement

Our study involves expert annotation and evaluation of fossil images, as well as user studies with citizen scientists evaluating AI-assisted fossil identification and

image-capture interfaces. Participation was voluntary, and no sensitive personal information or personal attributes were collected. The studies were conducted solely to evaluate fossil annotation quality, VLM capabilities, image-quality guidance, and interaction design for AI-assisted fossil identification.

In this work, we used generative AI tools as models evaluated in our experiments. Generative AI tools were not used to generate or modify ground-truth annotations; all annotations were provided by domain experts.

Reproducibility statement

We release the source code and materials supporting the three components of this work: (1) anonymous.4open.science/r/annotation_framework/README.md, (2) anonymous.4open.science/r/fossil_image_quality/README.md, and (3) anonymous.4open.science/r/interface/README.md. These resources include the data-processing and analysis code, experimental procedures, and evaluation materials required to reproduce the reported results. Further methodological and experimental details are provided in the corresponding sections and appendices.

References

1. Anbalagan, S.K., Nie, X., Mohan, U., Kanamarlapudi, V.K., Kommalapati, A., Zhao, X.: Bridging UI design and chatbot interactions: Applying form-based principles to conversational agents. arXiv (2025). https://doi.org/10.48550/arXiv.2507.01862
2. Anthropic: The claude 3 model family: Model card and technical report. Tech. rep., Anthropic (2024), https://www.anthropic.com/news/claude-3-5-sonnet
3. Beck, J., Eckman, S., Kern, C., Kreuter, F.: Bias in the loop: How humans evaluate ai-generated suggestions. Harvard Data Science Review **8**(2) (2026)
4. Bent Lindow: Nhmd vertebrate paleontology collection (2026). https://doi.org/10.15468/YENBN7, https://www.gbif.org/dataset/8e1b54db-ce7f-402e-a36e-c1382f0b7b97
5. Brooks, J.L.: Counterbalancing for serial order carryover effects in experimental condition orders. Psychological Methods **17**(4), 600–614 (2012). https://doi.org/10.1037/a0029310
6. Burgess, J., Nirschl, J.J., Bravo-Sánchez, L., Lozano, A., Gupte, S.R., Galaz-Montoya, J.G., Zhang, Y., Su, Y., Bhowmik, D., Coman, Z., et al.: Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19553–19564. IEEE (2025)
7. Chen, P., Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., Duan, H., Huang, Z., Su, Y., et al.: Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. Advances in Neural Information Processing Systems **37**, 94327–94427 (2024)
8. Cunha, L.F., Yu, N., Silvano, P., Campos, R., Jorge, A.: Leveraging llms to improve human annotation efficiency with inception. In: European Conference on Information Retrieval. pp. 53–58. Springer (2025)

9. Eijkelboom, I., Hogeweg, L., Brunink, D., Jansen, M., Ahrens, H., Schouwenburg, C., ter Haar, F., den Ouden, N., Langeveld, B.W., Amkreutz, L., et al.: Ai, citizen science data and expert opinions: towards an improved fossil identification process. PeerJ Computer Science **12**(e4049) (2026)
10. Eijkelboom, I., Schulp, A.S., Amkreutz, L., Verheul, D., Verschoof-van der Vaart, W., Van der Vaart-Verschoof, S., Hogeweg, L., Brunink, D., Mol, D., Peeters, H., et al.: Making sense of fossils and artefacts: a review of best practices for the design of a successful workflow for machine learning-assisted citizen science projects. PeerJ **13**, e18927 (2025)
11. Fadhil, A.: Domain specific design patterns: Designing for conversational user interfaces. arXiv (2018). https://doi.org/10.48550/arXiv.1802.09055
12. Google: Gemini 2.5 flash-lite. https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite (2025), accessed: 2026-06-15
13. Gu, Y., Fu, J., Liu, X., Valanarasu, J.M.J., Codella, N.C., Tan, R., Liu, Q., Jin, Y., Zhang, S., Wang, J., et al.: The illusion of readiness in health ai. arXiv preprint arXiv:2509.18234 (2025)
14. Hasinoff, S., Sharlet, D., Geiss, R., Adams, A., Barron, J., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Transactions on Graphics **35**, 1–12 (11 2016). https://doi.org/10.1145/2980179.2980254
15. He, G., Aishwarya, N., Gadiraju, U.: Is conversational XAI all you need? Human-AI decision making with a conversational XAI assistant. In: Proceedings of the 2025 International Conference on Intelligent User Interfaces. pp. 907–924. Association for Computing Machinery (2025). https://doi.org/10.1145/3708359.3712133
16. International Telecommunication Union: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. International Telecommunication Union, Geneva, Switzerland (2011), https://www.itu.int/rec/R-REC-BT.601/, recommendation ITU-R BT.601-7
17. Kevin Holston: Palaeozoological collections (pz), swedish museum of natural history (nrm) (2026). https://doi.org/10.15468/ZFFOLJ, https://www.gbif.org/dataset/20ab8294-5253-46ca-bdb2-eab1b5e75657
18. Kim, S., Lee, J., Gweon, G.: Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12. Association for Computing Machinery (2019). https://doi.org/10.1145/3290605.3300316
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
20. Luo, L., Chen, H., Xiao, Y., Zhou, Y., Wang, X., Vardhanabhuti, V., Wu, M., Han, C., Liu, Z., Fang, X.H.B., et al.: Rethinking annotation granularity for overcoming shortcuts in deep learning–based radiograph diagnosis: A multicenter study. Radiology: Artificial Intelligence **4**(5), e210299 (2022)
21. López-Guillén, E., Herrera, I., Bensid, B., Gómez Bellver, C., Ibáñez, N., Jiménez-Mejías, P., Mairal, M., Mena-García, L., Nualart, N., Utjés-Mascó, M., López-Pujol, J.: Strengths and challenges of using inaturalist in plant research with focus on data quality. Diversity **16**, 42 (01 2024). https://doi.org/10.3390/d16010042
22. Maruf, M., Daw, A., Mehrab, K.S., Manogaran, H.B., Neog, A., Sawhney, M., Khurana, M., Balhoff, J.P., Bakış, Y., Altintas, B., et al.: Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. Advances in Neural Information Processing Systems **37**, 131035–131071 (2024)

23. Natasja den Ouden, Jeroen Creuwels: Naturalis biodiversity center (nl) - paleontology (2026). https://doi.org/10.15468/57F7N4, https://www.gbif.org/dataset/ea233ece-c4fb-4fb8-bf98-f16235c4144c
24. National Museum Of Natural History, B.A.O.S.: Fossil and subfossil mammal collection (2025). https://doi.org/10.15468/Z3KDF3, https://www.gbif.org/dataset/b480eb15-6af9-48af-8104-f6dd575d8452
25. Naturalis Biodiversity Center (NL): Paleontology. https://www.gbif.org/dataset/ea233ece-c4fb-4fb8-bf98-f16235c4144c (2026), accessed via GBIF.org on 2026-06-15
26. Nielsen, J.: Usability Engineering. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1994)
27. Norman, D.: The Design of Everyday Things. Basic Books (2013)
28. Oervondstchecker: Oervondstchecker. https://www.oervondstchecker.nl/, accessed: 2026-08-19
29. OpenAI: Hello gpt-4o: Openai's multimodal foundation model. OpenAI Technical Report (2024), https://openai.com/index/hello-gpt-4o/
30. Philipp, L., Liu, H., Grbic, S., Hering, A., Georgescu, B.: Organ masks applied in feature space improve weakly supervised scan-level ct classification. Scientific Reports (2026)
31. Portela, M.: Interfacing participation in citizen science projects with conversational agents. Human Computation **8**(2), 33–53 (2021)
32. Qwen Team: Qwen2-vl: To see the world more clearly. arXiv preprint arXiv:2409.12186 (2024)
33. Scott Rufolo, Kieran Shepherd, Margaret Currie: Canadian museum of nature fossil vertebrate collection (2026). https://doi.org/10.15468/BLXVML, https://www.gbif.org/dataset/e9c27954-85d1-4a87-bce4-1715bd606268
34. Skarlatidou, A., Hamilton, A., Vitos, M., Haklay, M.: What do volunteers want from citizen science technologies? A systematic literature review and best practice guidelines. Journal of Science Communication **18**(01), A02 (2019). https://doi.org/10.22323/2.18010202
35. Soni, H., Ivanova, J., Wilczewski, H., Bailey, A., Ong, T., Narma, A., Bunnell, B.E., Welch, B.M.: Virtual conversational agents versus online forms: Patient experience and preferences for health data collection. Frontiers in Digital Health **4**, 954069 (2022). https://doi.org/10.3389/fdgth.2022.954069
36. SPOTTERON: Fossilfinder. https://www.spotteron.com/fossilfinder/, accessed: 2026-08-19
37. Van Der Wal, R., Sharma, N., Mellish, C., Robinson, A., Siddharthan, A.: The role of automated feedback in training and retaining biological recorders for citizen science. Conservation Biology **30**(3), 550–561 (2016). https://doi.org/10.1111/cobi.12705
38. Vogt, L., Baum, R.: Using named graphs and knowledge graph template patterns for efficiently organizing fair anatomy data and metadata. Biodiversity Information Science and Standards **3**, e37205 (2019)
39. Wang, W., Zhang, G., Li, M., Wang, Y., et al.: Coralvqa: A large-scale visual question answering dataset for coral reef image understanding. Advances in Neural Information Processing Systems **38** (2026)
40. Xie, C., Wang, B., Kong, F., Li, J., Liang, D., Zhang, G., Leng, D., Yin, Y.: Fg-clip: Fine-grained visual and textual alignment. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 68777–68793. PMLR (2025)

41. Yao, Y., Han, T., Li, D.: Assistance or distraction? a cognitive ergonomics perspective on cognitive load during ai-assisted post-editing. International Journal of Human–Computer Interaction pp. 1–20 (2026)
42. Yaqoob, M., Ishaq, M., Ansari, M.Y., Qaiser, Y., Hussain, R., Rabbani, H.S., Garwood, R.J., Seers, T.D.: Advancing paleontology: a survey on deep learning methodologies in fossil image analysis. Artificial Intelligence Review (2024). https://doi.org/https://doi.org/10.1007/s10462-024-11080-y, https://link.springer.com/article/10.1007/s10462-024-11080-y
43. Yu, C., Qin, F., Watanabe, A., Yao, W., Li, Y., Qin, Z., Liu, Y., Wang, H., Jiangzuo, Q., Hsiang, A.Y., Ma, C., Rayfield, E., Benton, M.J., Xu, X.: Artificial intelligence in paleontology. Earth-Science Reviews **252**, 104765 (2024). https://doi.org/https://doi.org/10.1016/j.earscirev.2024.104765, https://www.sciencedirect.com/science/article/pii/S0012825224000928
44. Zhang, Z., Ma, X., Chowdhury, A., Campolongo, E., Thompson, M., Zhang, N., Stevens, S., Lapp, H., Berger-Wolf, T., Su, Y., et al.: Biocap: exploiting synthetic captions beyond labels in biological foundation models. In: International Conference on Learning Representations. vol. 2026, pp. 125415–125451 (2026)
45. Zhou, Y., Faith, T.L., Xu, Y., Leng, S., Xu, X., Liu, Y., Goh, R.S.: Benchx: A unified benchmark framework for medical vision-language pretraining on chest x-rays. Advances in Neural Information Processing Systems **37**, 6625–6647 (2024)

A Fossil identification challenge

Unlike modern biology identification, typically involving observing complete living organisms, paleontological remains (typically from Pleistocene, a geological epoch that lasted from approximately 2.58 million to 11,700 years ago) are affected by a *taphonomic* filter. Specimens are frequently reduced to isolated fragments, heavily abraded by water transport and mineralized or stained by anaerobic sediments (see Figure 8). This process often obscures the very diagnostic features, such as fine enamel folds or nutrient, required for species-level identification. The difficulty is compounded by morphological convergence, where the skeletal elements of distantly related taxa evolve similar shapes due to shared locomotor or dietary requirements. For example, the tarsal and carpal bones (ankles and wrists) of large artiodactyls (e.g., bison vs. deer) exhibit a high degree of similarity when subjected to the rounding effects of marine reworking. To a non-expert, a fragmented "black bone" from a modern domestic cow can appear indistinguishable from that of a late Pleistocene steppe bison. Nevertheless,

Fig. 8: **Fossil identification challenge**. All fossils evidence of femur fragments belonging to order *Carnivora*. Source: Oervondstchecker.nl [28].



Fig. 9: **Examples of four different femur fragments of *Coelodonta antiquitatis* (woolly rhinoceros) from standardized collection, illustrating variation in appearance across specimens (vertical) and viewpoints (horizontal)**. Source: Naturalis Biodiversity Center (NL) - Paleontology [23].



the collection and expert annotation of fossil data are essential for reconstructing past ecosystems, understanding evolutionary processes, and revealing how species and environments have changed over time.

Citizen science. Citizen science has become an increasingly important approach for collecting large volumes of scientific data across diverse domains. These include ecology, biodiversity monitoring, astronomy, and paleontology. By enabling non-experts to contribute observations, citizen science projects can expand and accelerate data collection in addition to fostering public engagement with scientific research. However, the scientific value of citizen-contributed data is highly dependent on its quality, especially when observations are submitted in the form of photographs [21]. Image-based submissions are now included in many citizen science platforms, because they allow experts and automated systems to validate observations remotely. In fields such as paleontology, where physical access to specimens is often limited, photographs serve as the primary medium for identification, verification, and archival documentation [42]. Figure 8 shows five femur fragments belonging to order *Carnivora* recorded by citizen scientists in Oervondstchecker.nl [28], for which identification beyond the order taxonomic level is not possible based on the available evidence. Poor image quality can substantially reduce the usefulness of submissions, increasing expert workload, lowering identification accuracy, and limiting the effectiveness of machine learning models trained on such data [9].

Citizen-science technologies therefore need to balance data quality with accessible participation. Previous research highlights the importance of clear data-entry flows, feedback, learning opportunities, and community-oriented design in maintaining volunteer engagement and supporting reliable contributions [34,37].

Table 6: **Existing fossil datasets.** GoldenFossils (ours) extends in annotation richness (see Table 8 and Table 9).

Type	Dataset	Images	Annotations
standardized collection	Naturalis Biodiversity Center (NL) - Paleontology [23]	86.2K	taxonomic/anatomic labels
	NHMD Vertebrate Paleontology Collection [4]	241	taxonomic label
	Canadian Museum of Nature Fossil Vertebrate Collection [33]	919	taxonomic/anatomic labels
	Fossil and Subfossil Mammal Collection [24]	6.1K	taxonomic/anatomic labels
	Palaeozoological Collections (PZ), Swedish Museum of Natural History (NRM) [17]	5.4K	taxonomic labels
citizen-science collection	Oervondstchecker.nl [28]	75K	discoverer's note, partial expert validated taxonomic/anatomic labels
	Fossilfinder [36]	568	discoverer's note, partial expert validated taxonomic/anatomic labels

AI-assisted identification. Accurate identification often depends on subtle morphological features that need good lighting, sufficient sharpness, clear contrast between specimen and background, the inclusion of scale references, and multiple, relevant viewing angles [10]. As a result, variations in photographic quality can directly affect whether a submission can be meaningfully evaluated by experts or automated systems.

Current state-of-the-art models achieve only 50% accuracy trained on >150K image-label samples from community science and standardized collections (see Table 6: Oervondstchecker.nl [28] and Naturalis Biodiversity Center (NL) - Paleontology [25]), with substantially higher performance on standardized museum images than on heterogeneous citizen-science images [9]. This limitation stems from both the quality of fossil images and the limited contextual information available for identification. Examples of image quality of femur fragments from community science and standardized collections are shown in Figure 8 and Figure 9, respectively. Fossil occurrences are typically indexed by taxonomic identity, location, and time, but lack detailed morphological traits and various view angles. Furthermore, richer descriptors (e.g., texture, color, size) in community-science platforms such as Oervondstchecker.nl [28] and FossilFinder [36] are optional, inconsistently applied, or often absent altogether. Recent work in the context of AI-assisted identification (for modern biodiversity) demonstrates that incorporating contextual information and richer descriptions can enable models to achieve performance comparable to substantially larger-scale models [44].

Research aim. Our aim is therefore to improve fossil data collection at its source and enable more reliable downstream AI-assisted identification. Rather than treating identification as a one-step image-to-label task, we investigate how AI can support the complete data-collection and identification process: guiding citizens to capture higher-quality images, eliciting relevant information through

accessible interactions, and supporting experts in creating richer, localized morphological annotations. Better images and richer annotations can provide both experts and AI models with more informative evidence for identification, while generating higher-quality data for future model development.

We investigate three complementary directions toward this goal. First, we introduce GoldenFossils, a proof-of-concept dataset with localized, fine-grained morphological annotations created through an AI-assisted human-in-the-loop workflow and evaluate state-of-the-art VLMs to determine how richer, localized information affects their fossil identification and reasoning capabilities. Second, we investigate conversational and structured interfaces for guiding citizen scientists through fossil identification and data collection, examining their usability, efficiency, and user preferences [34,37,31]. Third, we investigate AI-assisted image-quality guidance at the point of capture, comparing no feedback, post-capture feedback, and real-time feedback for improving lighting, sharpness, contrast, scale, and viewing angles.

Together, these studies shift the focus from applying AI to existing fossil images toward using AI to improve the quality of the data on which identification itself depends, creating a feedback loop in which better AI assistance enables better data collection, which in turn supports more reliable AI-assisted fossil identification.

Use case. We focus specifically on vertebrate fossils, where identification often depends on fine-grained anatomical and morphological characteristics.

B GoldenFossils dataset

Unlike traditional datasets, which often lack rich textual descriptions beyond basic labels (see Table 6), GoldenFossils is a high-fidelity, expert-steered AI documentation. It captures not only the final taxonomic conclusion but also the iterative refinement process and the specific morphological justifications provided by domain experts (N=10).

The dataset generated during the evaluation phase consists of a taxonomically diverse array of fossil fragments, primarily focused on North Sea specimen (e.g. *Mammuthus primigenius*) from the Dutch region. The data was uploaded by citizen scientists to the Oervondstchecker.nl [28]. Each entry contains a complete scientific profile including stratigraphic context, mineralization states, and precise anatomical identifications. These fields represent the combined output of the VLM's initial hypothesis and the expert's subsequent corrections. A distinguishing feature of this dataset is the inclusion of coordinate-based image annotations. Each specimen is accompanied by localized markers that pinpoint the exact visual evidence used to reach the scientific conclusion. These specimens can be observed in the Figure 10. Experts have initially annotated these specimen which can be seen in Table 7, due to the size of each entry this table only shows the scientific name that was given as well as the description of the preserved part. A preview of a full entry is visible in Table 9.



Fig. 10: **Visual data sampled from Oervondstchecker.nl [28] (N=20), included in GoldenFossils.** Some fossils have multiple views image evidence. Table 7 illustrates the annotations associated with these specimens.

Licensing. To promote open science and foster future developments in automated paleontological identification, the GoldenFossils dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Data collection

Domain experts (other than the ones who initially validated the source fossil evidence) utilized the AI-assisted annotation framework (see Appendix F) an operational environment to document raw findings from a predetermined dataset. For this study, ten domain experts (researchers and scientist from Utrecht University's Stratigraphy & Paleontology group and Naturalis Biodiversity Center) were specifically tasked with annotating the specimen images, a process that captured not only the final taxonomic conclusions but also the iterative refine-

Table 7: **Fossil data sampled from Oervondstchecker.nl [28] (N=20).** These samples are annotated by experts with our novel AI-assisted annotation framework and included in the final GoldenFossils dataset. Figure 10 illustrates the figures associated with these annotations.

Ref	Scientific name	Annotation (Oervondstchecker.nl [28])
(a)	*Cervus elaphus*	Distal metacarpal (Red Deer). Note the "double rollers".
(b)	*Cervus elaphus*	Astragalus of a ruminant artiodactyl. Transitions smoothly.
(c)	*M. primigenius*	Lamelle of a molar. Thin and wide proportions.
(d)	*M. primigenius*	Deciduous third premolar (dp3).
(e)	*Bovidae*	Partial right mandible. Likely domestic cattle.
(f)	Not Identifiable	Large land mammal bone fragment.
(g)	*M. primigenius*	Heavily worn down "spit tooth".
(h)	*M. primigenius*	Jugal bone (cheek bone) with large trabeculae.
(i)	Not Identifiable	Indeterminate bone fragment (not antler).
(j)	*Equus ferus*	Upper M3 (third molar) of a horse.
(k)	*M. primigenius*	Root fragment of a mammoth tooth.
(l)	Not Identifiable	Indeterminate bone fragment.
(m)	*M. primigenius*	Right mandible of a young mammoth.
(n)	*Cervus elaphus*	Astragalus (Red Deer size).
(o)	*Cervidae indet.*	Partial metacarpal. Slender build rules out Bovidae.
(p)	*Bovidae indet.*	Scaphoid (wrist bone) of a ruminant.
(q)	Not Identifiable	Likely m3 of a pig (needs occlusal view).
(r)	*Cervidae indet.*	Distal epiphysis of an artiodactyl metapodial.
(s)	*Equus ferus*	"Tooth cap" (deciduous premolar remnant).
(t)	*Equus ferus*	p3 or p4 from the lower jaw.

Table 8: **GoldenFossils annotations of distal metacarpal of *Cervus elaphus* (red deer).** See visual evidence in Figure 10(a).

Annotation (Oervondstchecker.nl [28])	Annotation (ours)
Distal metacarpal (Red Deer). Note the "double rollers".	Distal metacarpal; the shape of this bone is very typical of artiodactyls (double rollers) and thus easily recognised at this level. Of note is that a metatarsal is distinguished because it would have a deep groove running along the anterior surface. Reindeer can be distinguished because the distal joint flares out significantly more than in the fossil. The distal width of ca. 45 mm best fits C. elaphus; fallow deer are significantly smaller, while moose and Irish elk are larger.

ment and specific morphological justifications provided by the expert group. This predetermined dataset is built upon random data entries from Oervondstchecker.nl [28], which is a prominent citizen science portal that allows users to upload their fossil findings and have them appraised by real domain experts. These entries consist of an image and small note to the future reviewer. However, these entries typically lack structured morphological metadata. The structure of the source dataset and the resulting dataset can be seen in Table 9. The source dataset component is defined by primary fields including the note submitted by the finder on Oervondstchecker.nl [28], and an array of links that point to the images belonging to that specific submission. The remaining components of the architecture represent the new GoldenFossils dataset. These elements include all image-related data, such as the annotations, as well as the scientific documentation data, which encompasses the collection category, geological period,

Table 9: **GoldenFossils dataset card, detailing administrative, stratigraphic, anatomical, and machine-actionable fields derived from the relational schema.**

Category	Field descriptions
Metadata	Unique identifiers (`id`, `expert_id`), collection naming, and creation timestamps (`created_at`) for version control and auditing; `basis_of_record` and `collection_name` define the nature of the entry (e.g., physical specimen vs. cast) and its institutional location.
Scientific identity	`scientific_name` using binomial nomenclature, supported by `geologicalPeriod` for broad temporal categorization.
Stratigraphic context	Multi-layered geological data including `chrono_strat`, `litho_strat`, and `bio_strat` defining the specific find-site formation.
Taphonomy	Combined physical state data: `preservation` quality, `condition_death` (state at time of burial), and taphonomic wear.
Anatomy	`preserved_part` and `anatomical_description` identifying specific skeletal elements (e.g., *tibia*) or morphological fragments.
Physical metrics	Expert-validated quantitative dimensions (`length_mm`, `width_mm`, `thickness_mm`) recorded as high-precision numeric values.
Image annotations	`image_url` and `annotations` (JSONB); includes coordinate-based visual justifications linking morphology to specific pixels.
Provenance	Links to the `discovery_queue` via `discoveryKey`. Includes `discoverer_note` (original field observations) and `note_to_finder`.
Processing state	Status flags including `annotated` and `gettingEdited` (boolean) to track the manual-to-machine validation pipeline.

and scientific name. Combined, these components provide a machine-actionable repository of expert justifications that can be used for future model fine-tuning [44].

C Dataset entry record preview

Discovery ID:	6
Database UUID:	`0f324bbf-c53e-418c-830d-4bbb5aa9da78`
Scientific name:	*Cervus elaphus* (red deer/edelhert)
Geological period:	Midden Weichselien / Vroeg Weichselien 20.000 – 65.000
Collection:	Fossil Vertebrates
Basis of Record:	Fossil Specimen
Preservation:	Permineralisation
Width (mm):	45
Anatomical part:	Distal metacarpal; the shape of this bone is very typical of artiodactyls (double rollers) and thus easily recognised at this level. Of note is that a metatarsal is distinguished because it would have a deep groove running along the anterior surface. Reindeer can be distinguished because the distal joint flares out significantly more than in the fossil. The distal width of ca. 45 mm best fits *C. elaphus*; fallow deer are significantly smaller, while moose and Irish elk are larger.
Note to Finder:	Beste vinder, dit is het onderste gewricht van een middenhandsbeen (middenvoetsbeen klopte dus bijna!) van een edelhert.

Image URLs:	`.../specimen_0025.jpg`
	`.../specimen_0026.jpg`
	`.../specimen_0027.jpg`
	`.../specimen_0028.jpg`

Annotations:	No annotations recorded for this entry.

D Image quality

Paleontological citizen science has distinct challenges for image-based data collection. Fossils are often fragmentary, partially embedded in matrix material, and visually similar to their surrounding background. Accurate identification often depends on subtle morphological features that need good lighting, sufficient sharpness, clear contrast between specimen and background, the inclusion of scale references, and multiple, relevant viewing angles [10]. As a result, variations in photographic quality can directly affect whether a submission can be meaningfully evaluated by experts or automated systems. The current submission workflow, implemented through platforms such as Oervondstchecker.nl [28], allows users to upload photographs and metadata for expert review. While this approach is accessible and scalable, it places the responsibility for image quality



(a) Citizen science submission from Oervondstchecker.nl [28]

(b) Archived fossil from Naturalis Biodiversity Center (NL) - Paleontology [23]

Fig. 11: **Comparison between a citizen science fossil submission and an archived museum specimen.**

entirely on the contributor and does not guide the image capture process. Consequently, quality control occurs only after submission, often requiring follow-up communication or resulting in unusable data.

Figure 11 illustrates the difference between a typical citizen science fossil submission and a professionally archived museum specimen. Figure 11a shows an example of a fossil image that was taken by a citizen scientist. Figure 11b shows a fossil image from a paleontological archive at Naturalis Biodiversity Center [23].

While previous work in paleontological citizen science has concentrated on improving classification models or refining expert review workflows, little attention has been given to how contributors might be supported during the act of taking photographs itself. Improving image quality at the point of capture has the potential to reduce expert workload, increase the scientific usefulness of citizen-generated data, and enhance downstream machine learning pipelines. Yet it remains unclear whether real-time guidance, post-capture feedback, or the absence of feedback produces meaningfully different outcomes in terms of image quality or user experience.

D.1 Measuring image quality

The objective was to create a method for measuring image quality that is both objective and reproducible. To achieve this, a quantitative scoring system was developed to calculate image metrics based on an experiment conducted by an experienced paleontologist. The system incorporates the expert's experience and includes subjective grading. This grading combines both quantitative and qualitative assessments.

The grading system is based on a dataset of 100 fossil images submitted by citizen scientists. These images were evaluated by a paleontologist using a scale from one (very poor) to five (very good). The criteria for grading included

lighting, sharpness, and the contrast between the fossil and its background. Additionally, the evaluation considered the presence of a scale, the type of scale shown in the image, and whether the fossil is depicted from multiple, relevant angles.

To convert the expert's qualitative judgments into actionable decision rules, numerical thresholds were established for each metric. In the first step, the continuous values for lighting, sharpness, and contrast were matched with the expert's 1–5 ratings to create initial threshold candidates for each quality level. In the second step, these thresholds were iteratively refined through in-app testing on actual devices, ensuring that the boundaries between rating levels were both empirically sound and aligned with practical image capture conditions in the field.

For measuring image quality, three metrics can be calculated without a reference image. Lighting is computed as the mean gray value over all pixels in the image. The RGB values are first converted to grayscale using standard luminance weights ($0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B$) [16], and then the arithmetic mean across all pixels is calculated. Based on empirical testing, slightly stricter thresholds were defined to more reliably detect under- and overexposed images, with breakpoints at gray values 60, 90, 120, and 180.

Sharpness is measured using the variance of the Laplacian in the central region of the image. A discrete Laplacian operator is applied to the middle 50% of the image area, and the variance of the resulting Laplacian values is computed. Higher variance corresponds to stronger edges and, therefore, better focus. The thresholds 40, 80, 110, and 150 were calibrated using real device data so that slightly shaky or soft images are treated more leniently, while clearly focused images are consistently classified as "good" or "very good."

Contrast between the fossil and its background is captured using a center–edge contrast measure. In the grayscale image, a central circular area (about 60% of the image diameter) is defined as the object region, and a more distant ring is treated as background. The mean brightness of the center region and the edge region is computed, and the absolute difference between them is used as the contrast metric. Higher values indicate stronger separation between the specimen and the background. Thresholds of 15, 30, 40, and 60 impose a relatively strict standard on contrast, so that poorly separated or "blending" fossils are more clearly downgraded.

All three metrics are classified into five qualitative rating levels: "Very Poor," "Poor," "Intermediate," "Good," and "Very Good." A general rating function assigns a continuous metric value to one of these categories by using four thresholds. Values below the first threshold are labeled "Very Poor," those between the first and second thresholds are labeled "Poor," and so on. Values above the highest threshold are labeled "Very Good." The remaining qualitative aspects, such as the presence and type of scale and the relevance of the viewing angle, are manually evaluated and mapped to the same rating levels after data collection concluded.

D.2 Image quality benchmark

The collected images are used to assess the impact of the guidance and feedback on image quality using the image quality rating system described above. Testers followed the same process:

1. Taking one or more images.
2. Reviewing the provided feedback mechanism.
3. Being able to delete any unwanted images before submission.
4. Submitting their final selection of images.

A prototype application is the main tool for data collection. Its design and implementation followed user-centered design principles and an iterative approach, comparing no feedback, post-capture feedback, and real-time feedback showed that real-time guidance. Architectural decisions improve usability and technical performance. The design follows established usability heuristics, emphasizing clarity, consistency, and intuitiveness [26]. Existing applications, such as Google Lens and Rock Identifier, were analyzed to create an intuitive and familiar user experience. The visual styling is intentionally minimalistic. A natural color palette featuring green and beige tones emphasizes the paleontological context. Using system fonts ensures consistent readability across various devices. These choices are designed to minimize distraction and visual fatigue, allowing users to focus on capturing high-quality images. Figure 4 presents the interface of the prototype. The instructions provided for the baseline are the same as those found on the platform Oervondstchecker [28]. Additionally, it features three icons that represent lighting, sharpness, and contrast. These icons change color to indicate the quality: red (very poor or poor), yellow (intermediate), and green (good or very good). The scores are determined by the image quality rating system described above. When any of the icons are pressed, users receive a more detailed explanation of what each one represents, as well as suggestions for adding a scale and taking multiple angles.

Image quality was assessed using five metrics: **lighting**, **sharpness**, **contrast**, **scale rating**, and **angle rating**. Next, the calculated metrics are compared and their statistical significance is determined.

Objective quality metrics were analyzed to determine whether feedback timing influenced the technical quality of submitted images. Descriptive statistics for each metric are presented in Table 1, showing the mean, standard deviation, and median values for lighting, sharpness, contrast, scale rating, and angle rating across each prototype condition. The Real-Time Feedback prototype demonstrated a higher median contrast score (41.88) compared to the Baseline (18.63) and Post-Capture Feedback (31.85) prototypes, suggesting a potential improvement in this dimension. Kruskal-Wallis H-tests were conducted for each metric to assess the statistical significance of differences across prototype conditions. A significant difference was observed only for the contrast metric ($H = 9.03$, $p < \alpha$). Pairwise comparisons using Mann-Whitney U tests with Bonferroni correction ($\alpha = 0.025$) revealed that the difference in contrast between the Baseline and Real-Time Feedback prototypes was statistically significant ($U = 1236.00$, $p < \alpha$). The median contrast score for the Real-Time Feedback proto-

type was 125% higher than the Baseline, indicating substantial improvement in object-background separation. No significant differences were found for lighting, sharpness, scale rating, or angle rating across the three prototype conditions.

D.3 Limitations and potential sources of bias

Several limitations and potential sources of bias should be considered when interpreting the findings of this study. First, the within-subject experimental design required all participants to evaluate all three prototypes. While this design enabled direct comparison between feedback conditions and reduced inter-participant variability, it may also have introduced comparative bias. Participants were aware that they were evaluating multiple versions of the same system, which may have encouraged relative rather than absolute judgments, particularly in usability ratings.

Second, the intended balanced randomization of prototype order failed during the final data collection phase, resulting in the same testing sequence for 82,6% of participants (Baseline → Real-Time Feedback → Post-Capture Feedback). This ordering may have introduced learning effects or contrast effects between conditions. Although the real-time feedback prototype was evaluated before the post-capture prototype, prior exposure to feedback mechanisms may still have influenced subsequent interactions.

Third, participants were aware that they were participating in a research study, which may have influenced their behavior and attentiveness (Hawthorne effect). Finally, the evaluation focused on short-term interactions with prototype systems rather than long-term use, limiting insight into learning effects and long-term behavioral change.

D.4 Future work

Several directions for future research follow directly from the limitations of this study. First, the experimental design can be improved by introducing randomized or counterbalanced ordering of the prototypes. This would reduce learning and order effects and allow a clearer separation of the impact of feedback timing from participant familiarity with the task.

Second, the long-term effects of image quality feedback are still unexplored. Future work should investigate whether repeated exposure to real-time feedback leads to sustained improvements in image quality or learning effects over time. Longitudinal studies could assess whether users internalize quality guidelines and need less guidance as experience increases.

Third, the feedback mechanisms themselves can be refined. The current system relies mainly on textual and visual instructions. Future versions could incorporate fossil image examples, other overlays, or adaptive feedback that responds to repeated user errors. More explicit guidance for scale inclusion and viewing angle may be necessary, as these aspects did not improve in the current study.

Finally, the image quality assessment approach can be expanded. Additional features, such as automatic detection of scale objects, fossil orientation, or occlusion, may better reflect scientific usability. Combining these features with AI performance measures could help clarify the connections between capture guidance, image quality, and model accuracy. Together, these improvements would support a more robust evaluation of real-time feedback systems and their role in improving data quality within citizen science projects.

D.5 Conclusion

This study examined whether feedback on image quality can improve fossil photographs submitted by citizen scientists. Comparing no feedback, post-capture feedback, and real-time feedback showed that real-time guidance significantly improved fossil-background contrast and achieved higher usability than the baseline, while no significant differences were found for lighting, sharpness, scale inclusion, or viewing angle. Post-capture feedback produced no significant improvements, suggesting that feedback provided during image capture may be more effective than delayed guidance. Overall, the findings demonstrate the potential of real-time guidance to improve specific aspects of citizen-generated fossil data, while highlighting the need for more comprehensive feedback mechanisms and evaluation under real-world conditions.

E Conversational vs. form-based interface for AI-assisted fossil identification

Conversational interfaces have been proposed as a promising alternative to traditional web-based interfaces in citizen-science and data-collection contexts. Conversational agents may lower barriers to participation by providing a more familiar, guided, and socially engaging interaction style [31]. Empirical comparisons between chatbots and web forms suggest that conversational interfaces can increase engagement, enjoyment, or perceived interaction quality, but often at the cost of longer completion times [18,35]. Traditional forms, by contrast, provide visible structure, overview, and clear input constraints, which can reduce ambiguity and make data entry more predictable. This suggests a recurring trade-off between the efficiency and structure of form-based interfaces and the experiential qualities of conversational interaction.

At the same time, these benefits do not mean that conversational interfaces are inherently superior. Conversational systems can suffer from ambiguity, unclear correction mechanisms, and context-management problems, especially in multi-step task-oriented interactions [11]. Natural-language explanations may also make AI systems feel more complete or persuasive than they actually are, potentially increasing over-reliance [15]. Recent work therefore suggests that hybrid approaches, which combine conversational guidance with form-like structure, may be a promising direction for improving task coherence and reducing

user confusion [1]. However, designing such hybrid systems requires a clearer understanding of what users value in each interaction style separately, and whether the trade-offs reported in prior work also apply in the specific context of AI-assisted fossil identification.

We employed a within-subject comparative design to evaluate the short-term user experience of two interface paradigms for AI-supported fossil identification: **conversational chat interface** and **transactional form-based interface**. Both conditions were implemented within the same application, ensuring that differences between conditions were attributable primarily to interaction structure rather than to differences in backend logic or overall visual identity. Because each participant completed both conditions, the study enabled direct within-participant comparison of the two interface types.



Fig. 12: **Fossil cases used in the interface study.** Each case is represented by three standardized views: the first row shows the mammoth scapula case, and the second row shows the wild horse metatarsal case.

E.1 Data collection

A custom fossil-identification application to support citizen scientists through expert-like guidance was designed. Two interaction conditions were compared:

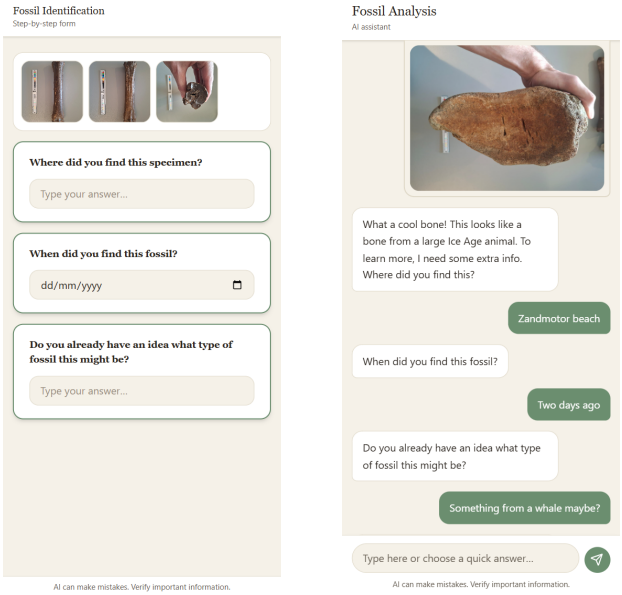


Fig. 13: **Representative screens from the two interface conditions.** The transactional condition presented the identification task as a structured step-by-step form with input fields and predefined options. The conversational condition presented the same task as a chat-style exchange with conversational prompts, user message bubbles, and quick-answer options.

(1) a conversational interface using chat-style prompts and quick-answer options, and (2) a transactional interface using structured, step-by-step form fields (see Figure 13). Both followed equivalent identification logic and produced standardized results containing the proposed identification, scientific name, geological period, confidence, and explanation. Participants completed both fossil-identification tasks, followed by post-task evaluations and a final comparative questionnaire.

Two scripted fossil-identification cases were used in the study flow. Each case was represented by three standardized images showing the fossil from different views (see Figure 12. The first case represented a large, flat fossil bone that led to the scripted identification of a woolly mammoth scapula (*Mammuthus primigenius*). The second case represented a long, robust fossil bone that led to the scripted identification of a wild horse metatarsal (*Equus caballus*).

The sample comprised 51 participants distributed across four age clusters. As shown in Table 2, young adults formed the largest age group, but the sample included participants across all four age clusters. Participants reported moderately high tech savviness, with a mean of approximately 5.4 on a 7-point scale, and moderate fossil expertise, with a mean of approximately 4.1. Overall, the sample therefore represented users who were generally comfortable with digital

systems and reported moderate levels of self-perceived domain expertise in fossil identification.

E.2 Results

Overall interface preferences are summarized in Table 10. When asked which interface they preferred overall, 27 participants (52.9%) selected the chat interface and 24 participants (47.1%) selected the form interface. A binomial test against an even 50–50 split was not significant ($p = 0.7798$), indicating that the sample did not show a clear aggregate preference for one interface.

Preference was broadly aligned with the average rating pattern, but not perfectly. In 35 cases (68.6%), participants preferred the same interface they had rated higher on average. In 5 cases (9.8%), participants preferred the interface they had rated lower. In the remaining 11 cases (21.6%), both interfaces received equal average ratings. Direct preference therefore overlapped substantially with the numeric evaluations, but it also captured considerations that were not fully reflected in mean rating differences alone.

Table 10: Overall interface preference counts, percentages, and preference-rating alignment categories.

Measure	**Category**	**Value**
Preferred interface	Chat	27 (52.9%)
Preferred interface	Form	24 (47.1%)
Binomial test	Chat vs. form	$p = 0.7798$
Preference-rating alignment	Match	35 (68.6%)
Preference-rating alignment	Mismatch	5 (9.8%)
Preference-rating alignment	Equal average rating	11 (21.6%)

E.3 Implications for AI-assisted fossil-identification tools

The findings suggest that future AI-assisted fossil-identification tools should not treat conversational and form-based interfaces as mutually exclusive alternatives. Instead, the results point toward the value of hybrid interaction designs. A purely form-based interface can provide faster completion times, structure, and overview, while a conversational interface can provide guidance, flexibility, and a more natural interaction experience. Recent work on conversational interface design similarly suggests that form-like principles, such as explicit confirmation, correction, and context switching, can be integrated into conversational systems to reduce ambiguity and improve task coherence [1]. For fossil identification, such a hybrid approach could combine structured fields for scientifically important metadata with conversational prompts that help users understand what to look for and why certain features matter.

This hybrid direction is also relevant for the broader citizen-science challenge of collecting useful data for machine learning. Paleontological AI remains constrained by limited high-quality labeled data [43], and citizen-science workflows

involving fossils and artifacts require careful attention to data quality, volunteer support, and iterative improvement [9]. Interface design is therefore not a secondary concern. If an interface is too rigid, it may reduce engagement and discourage continued use. If it is too conversational and unconstrained, it may increase ambiguity or reduce standardization. The most promising direction may be an interface that preserves the motivational and explanatory strengths of conversation while retaining the clarity and data structure needed for scientific use.

E.4 Limitations

Several limitations should be considered when interpreting the findings. First, the sample size was modest, which limits the statistical power of subgroup analyses and makes descriptive patterns, such as possible age-related differences, tentative rather than confirmatory. Second, the two fossil cases and two interface conditions were counterbalanced only within the constraints of the implemented study flow. Because fossil case order and interface order were linked, the strong order effects cannot be attributed to a single cause. Third, the qualitative analysis was exploratory and based on short open-text explanations rather than in-depth interviews.

E.5 Future work

Future research should investigate hybrid interface designs that combine structured data entry with conversational guidance. For example, such a hybrid interface could use structured fields for essential metadata, such as location and date while using conversational prompts to guide users through more uncertain observations such as shape, texture, weight, fossilization, and anatomical features. This would make it possible to test whether the benefits observed in both interfaces can be combined: the interaction speed and clarity of forms together with the guidance and engagement of conversation. Future work should also separate interface order from fossil-case order more fully, include larger and more diverse participant samples, and evaluate performance across a wider range of fossil-identification scenarios.

In addition, future studies should move beyond preference and completion time by examining data quality, learning, trust, and longer-term engagement. For citizen-science applications, the best interface is not necessarily the one users like most or complete fastest, but the one that supports sustained participation, accurate data collection, and appropriate interpretation of AI assistance. These outcomes are especially important in scientific domains involving fossils, artifacts, and other physical specimens, where high-quality labeled data remain a central bottleneck for machine learning development.

E.6 Conclusion

This study compared conversational and form-based interfaces for AI-assisted fossil identification. Both were evaluated positively, with no clear overall pref-

erence. Preferences reflected different strengths: forms offered structure, clarity, and efficiency, while chat provided guidance, flexibility, and engagement. However, strong order effects indicate that preferences were influenced by interaction sequence and should be interpreted cautiously. Overall, the findings motivate hybrid interfaces that combine structured data entry with conversational guidance, balancing efficiency, usability, data quality, and appropriately calibrated trust.

F AI-assisted annotation framework

The framework was developed to address the specific need for structured, machine-actionable data in paleontological research, moving away from traditional unstructured text notes. To ensure the framework's utility within a professional museum environment, the data schema and functional requirements were strictly modeled after the internal documentation standards of Naturalis fossil dataset [25]. This design choice ensures long-term interoperability with existing paleontological databases. The system's integrity is safeguarded by Supabase Authorization, which acts as a centralized gatekeeper ensuring that moderation and quality criteria meet established European standards for citizen science [36]. Access to the framework is strictly limited to authenticated personnel. Without a valid institutional user account, the system denies access to the underlying discovery queue and documentation tools. The front-end serves as the primary workstation for the domain expert, utilizing React 19 and React Router v7 to manage specimen-specific states. The architecture displayed in Figure 16 demonstrates the internal design of the framework.

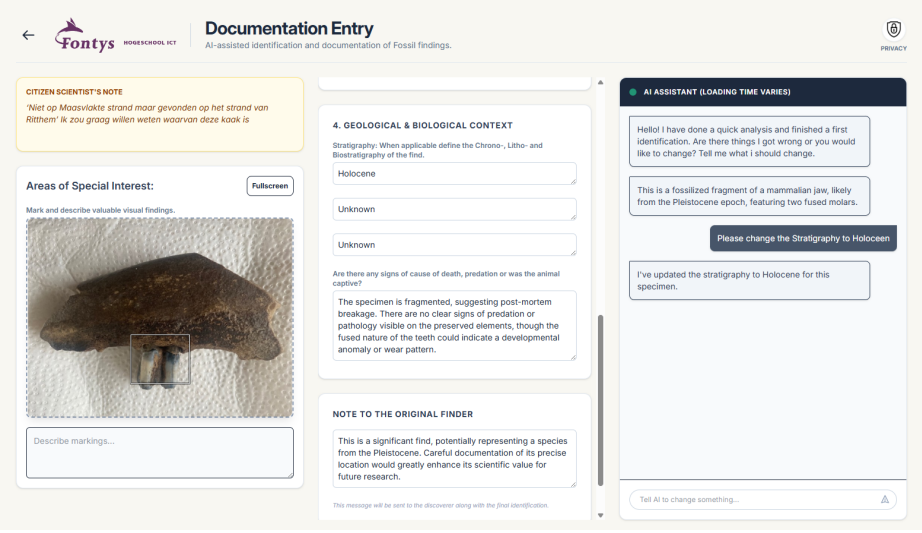


Fig. 14: **Screenshot of the central hub of the annotation framework.**

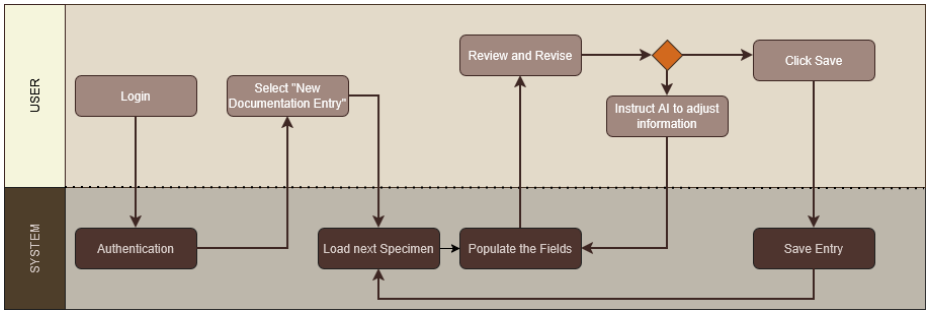


Fig. 15: **Simplified workflow diagram of the annotation framework.**

The UI model follows a streamlined, institutional workflow designed to reduce the cognitive burden on the expert while maintaining high-fidelity output. The workflow is divided into three primary stages: Upon successful authentication, users are redirected to the Central Hub which is displayed in Figure 14. This dashboard serves as the operational workstation's nerve center, providing the expert with two distinct pathways.

To create a new documentation entry or to view the archive. The archive allows experts to browse and filter previously completed entries for review and further refinement. When selecting to create a new entry, the framework automatically fetches the next priority specimen that requires labeling unlike general-purpose biodiversity tools that often focus on purely automated classification. When a specimen is loaded, the application utilizes the Google Gemini 2.5-flash-lite [12] model via the Generative AI SDK. The framework implements a pre-emptive analysis logic: as soon as the specimen's visual data is retrieved from Supabase Storage, the model performs an initial visual analysis.

This multimodal capability performs the "heavy lifting" by automatically populating the scientific fields, such as anatomical identification, preservation state, and geological period before the expert begins manual editing. This ensures that the expert transitions immediately into a review and refine role rather than a data entry role, significantly increasing documentation efficiency.

If the expert identifies nuances that the initial analysis overlooked, such as specific mineralization patterns or subtle taphonomic markers, they can utilize the conversational agent to steer the data. By providing natural language feedback, the expert can instruct the AI to update specific form fields in real-time, creating a collaborative human-AI documentation loop that results in an expert validated scientific record.

G Empirical evaluation

We benchmark current state-of-the-art VLMs capabilities in labeling and describing fossils over GoldenFossils.

Metrics. Taxonomic identification was quantified through Top-1 Accuracy, representing the mean match rate between the model's output and the expert-verified scientific name. To ensure statistical rigor, 95% Confidence Intervals (CI) were calculated and represented as error bars, illustrating the stability of model performance across the specimen dataset. Advanced benchmarking was conducted using mean Recall at k=10 (Recall@10) to establish standard Information Retrieval (IR) baselines, while mean Average Precision at k=50 (AP@50) was utilized to measure the ranked precision of the models. This calculation is particularly significant as it evaluates the reliability of the models' "intuition", rewarding instances where correct identifications were paired with high confidence scores. The distribution and shift of AI-reported certainty were analyzed using boxplots, focusing on the Interquartile Range (IQR) and median shifts between the two testing tiers. This method allowed for the identification of unearned confidence in the baseline tier and the quantification of increased certainty provided by crop assistance. Finally, the precision of the models' spatial reasoning and heavy lifting capabilities regarding scale were measured using Mean Absolute Error (MAE). This metric calculated the average absolute deviation in millimeters (mm) between the AI's estimated dimensions and the physical measurements recorded in the ground truth, providing a direct assessment of the models' remaining scaling deficits.

Models. The performance of this standard is measured against a representative sample of leading VLMs: Gemini 2.5 Flash [12] (framework's native engine) , GPT-4o [29], Qwen 3.6-Plus [32] and Claude 3.5 Sonnet [2]. To isolate how different levels of human assistance affect AI accuracy, the benchmarking is conducted across two distinct tiers. (1) **Test A: Zero-shot identification (baseline)** Models are provided with only the raw specimen image and the original discoverer's note. The models are prompted to identify the fossil and fill the established documentation schema. This tier measures the AI's autonomous baseline intuition. (2) **Test B: Localized feature analysis (crop-assisted)** In addition to the raw data, models are provided with the specific coordinates and visual crops of the areas of special interest defined by experts (see Appendix F). This evaluates whether human-led visual attention significantly improves taxonomic accuracy, a method previously explored in medical imaging but rarely in paleontology.

H More qualitative and quantitative evaluations over GoldenFossils

Test A: Zero-shot identification (baseline) In this setting, models were provided with a raw fossil image and a finder's note without additional guidance. We evaluate performance across $N = 20$ specimens (a–t), focusing on taxonomic identification, anatomical description, collection classification, and metric estimation. The models showed varying levels of success in identifying

both the scientific name and the anatomical element. Specimen-level predictions are summarized in Table 5.

Taxonomic performance was strongest for visually distinctive megafauna. All models correctly identified the *Woolly Mammoth (M. primigenius)* and *Wild Horse (Equus ferus)* when represented by characteristic dental material (c, t). However, the predictions also revealed a recurring *megafauna bias*, with models tending to favor prominent Pleistocene taxa when the visual evidence was ambiguous. Performance decreased substantially for post-cranial material. For example, specimen (a), a Red Deer metacarpal, was frequently misidentified as *Bison* or *Bos*, suggesting difficulty distinguishing subtler osteological characteristics.

More substantial failures were observed for ambiguous fragments. Specimen (i), an indeterminate bone fragment, was classified as wood by Gemini, ChatGPT, and Qwen. Similarly, Qwen classified specimen (j) as *Pinus* wood. All models also failed to identify the suspected pig tooth in specimen (q), instead proposing a range of unrelated Pleistocene taxa. These cases illustrate the difficulty of zero-shot identification when specimens lack distinctive or clearly visible diagnostic features.

For collection classification, model outputs were considerably more consistent, matching the ground truth in approximately 95% of cases. ChatGPT and Claude 3.5 Sonnet also produced more detailed preservation descriptions, frequently referring to processes such as "permineralization" and "anaerobic staining," which were broadly consistent with expert observations of iron and manganese oxide staining.

Metric estimation remained a major limitation. Models estimated specimen dimensions from visual cues alone, with estimated width compared against the available expert measurements. Larger fragments, such as specimens (f) and (h), were frequently underestimated by more than 50%. Estimates were somewhat more accurate for mid-sized specimens such as specimen (p), but overall the models could not reliably infer physical dimensions from unscaled images. Instead, they often appeared to rely on typical sizes associated with the predicted anatomical element rather than the actual scale of the photographed specimen.

Test B: Localized feature analysis (Crop-assisted) In this setting, models were provided with expert-cropped images focusing on diagnostic features (e.g., "linguaflexid," "double rollers," or "dp3 proportions"). This visual "pointing" reduced taxonomic error and increased the descriptive depth of model outputs. This resembles findings in medical imaging, where directed visual attention can improve diagnostic precision [7].

The consistency of metadata and the precision of size estimations were also evaluated. While ground-truth measurements were available for specific dimensions (e.g., width), models attempted to estimate all three axes. The "basis of record" field (see Appendix C) showed a high level of agreement with the ground truth, with 93.3% of responses correctly classifying specimens as "Fossil Specimen" or "Sub-Fossil."

Table 11: **Taxonomic identification performance under crop-assisted analysis using localized expert annotations and diagnostic visual features.** Correct identifications are highlighted in **bold**.

Key	Expert annotation	Gemini 3.5 Flash	GPT-4o	Claude 3.5 Sonnet
(b)	Straight, double rollers	***Cervus elaphus***	***Cervus elaphus***	*Capreolus capr.*
(c)	Overall thin and wide	*M. primigenius*	*M. primigenius*	*M. primigenius*
(d)	Proportion is a dp3	***M. primigenius***	***M. primigenius***	***M. primigenius***
(e)	High-crowned tooth	*Equus ferus*	*Equus ferus*	*Bos primigenius*
(g)	Remaining root	***M. primigenius***	*Cervus elaphus*	***M. primigenius***
(h)	Large trabeculae	***M. primigenius***	***M. primigenius***	***M. primigenius***
(j)	Enamel fold shape	***Equus ferus***	***Equus ferus***	***Equus ferus***
(k)	Striated surface texture	*Mammalia indet.*	*Mammalia indet.*	*Cetacea indet.*
(m)	Tooth lamelle imprints	***M. primigenius***	***M. primigenius***	***M. primigenius***
(n)	Smooth transition	***Cervus elaphus***	***Cervus elaphus***	***Cervus elaphus***
(o)	Division in the middle	*Mammalia indet.*	*Gallus gallus*	*Mammalia indet.*
(p)	Wide and deep (groove)	*Equus ferus*	*Bos primigenius*	*Bos primigenius*
(r)	Epiphyseal structure	*Bison priscus*	*Bos primigenius*	*C. antiquitatis*
(s)	Flat/Rootless tooth cap	*Cervus elaphus*	*M. primigenius*	*M. primigenius*
(t)	Linguaflexid/Ectoflexid	***Equus ferus***	***Equus ferus***	***Equus ferus***

Numerical measurements remained less reliable. Even when a scale bar was present, AI-generated measurements showed a mean deviation of 24% from the ground-truth values. For example, for specimen (c), models consistently overestimated the size of the molar lamella relative to the expert measurement (90 mm ground truth versus approximately 145 mm estimated). Standardized performance is summarized in Table 11.

The VLM's performance comparison Figure 18 (left) illustrates the mean taxonomic match rate for each model relative to the expert-verified ground truth. In Test A, the models achieved a mean accuracy ranging between 0.3 and 0.5. The vertical error bars indicate the 95% confidence interval, showing a high degree of variability in model performance across different specimen types.

In Test B (crop-assisted), a simultaneous increase in the mean match rate was observed across all four models. Accuracy scores in this tier reached a mean of 0.45. Additionally, the vertical error bars in Test B are shorter than those in Test A, representing a decrease in performance variability when the models were provided with localized morphological crops.

In Test A, the models showed a mean error of approximately 4.0486mm. GPT-4o and Gemini exhibited the highest deviations, with some estimates exceeding the ground truth by more than 10mm. In Test B, the point plot indicates a downward trend in measurement error for three out of the four models. While precision improved, the remaining error margin in the assisted tier maintained a mean of 3.157mm across the dataset. The data show that even with crop assistance, numerical estimations continued to deviate from the expert-recorded dimensions.

By isolating these failure points, we provide the empirical evidence needed to determine the reliability of human-AI collaboration in paleontological documentation compared to fully autonomous pipelines. Following the execution of these benchmarking tiers, the data generated by the VLMs will be cross-

referenced with the expert-verified ground truth. This comparative step is essential to highlight any performance gaps or technical limitations in the models' current visual reasoning. By identifying exactly where the AI's autonomous descriptions deviate from professional documentation, the research can pinpoint specific morphological features, such as intricate surface textures, mineralization states or rare taphonomic markers, where VLM performance is currently lacking. The Figure 19 shows a collage of samples that were consistently misidentified.

Discussion

The results of Test A reveal strong baseline competency for identifying distinctive Pleistocene taxa, alongside substantial morphological limitations and systematic biases. This section analyzes the factors influencing zero-shot performance and the discrepancies observed between model outputs and ground truth.

The models demonstrated near-perfect recall for visually distinctive specimens, such as *Mammuthus primigenius* molar lamellae (c, d, g, m) and *Equus ferus* hypsodont teeth (t). This success is likely attributable to the strong visual distinctiveness and prevalence of these morphologies in open-source paleontological imagery. The parallel enamel ridges of mammoth teeth and the complex folding patterns of equid molars provide salient visual features that can be readily associated with taxonomic labels.

However, performance decreased substantially for *Artiodactyla* specimens (a, b, n). While models often identified the broader group correctly, they struggled to distinguish between *Cervidae* and *Bovidae* without localized guidance. This suggests that current models possess a general morphological intuition but lack the expert-level sensitivity required for subtler osteological distinctions, such as differences in articular transitions and joint morphology.

One of the clearest findings in Test A was a pervasive megafauna bias. When presented with ambiguous fragments or common taxa, the models frequently defaulted to prominent Pleistocene species rather than maintaining uncertainty or selecting less visually salient alternatives.

Future work

While the current framework provides a technically sound environment for morphological justification, the volume of the resulting dataset remains insufficient for high-fidelity VLM fine-tuning. The expert gap observed during benchmarking indicates that human-in-the-loop validation provides substantially more reliable annotations than autonomous AI output, but the labor-intensive nature of expert annotation currently limits dataset scale. Future work should therefore focus on collaborative crowdsourcing across paleontological institutions and advanced students, expanding the dataset beyond the current pilot toward the originally targeted 2,000+ specimens. A larger dataset could support the development of a specialized Paleo-VLM adapter trained on the localized morphological features captured in this study.

A second important direction is to preserve the difference between the AI's initial hypothesis and the expert's final validation. The current framework prioritizes the final GoldenFossils record, which can obscure where and how the AI failed. Future versions should implement database versioning that automatically stores the initial VLM output as a permanent read-only record before expert editing. Such an audit trail would enable longitudinal analysis of recurring failure modes and model improvement across fossil categories.

Finally, the pre-analysis workflow could be extended with a visual feedback loop. When the AI updates a taxonomic or morphological field through dialogue, it should also highlight the image region that supports the revised conclusion. This would enable a more explicit bidirectional visual conversation and improve both transparency and trust in AI-assisted fossil identification.

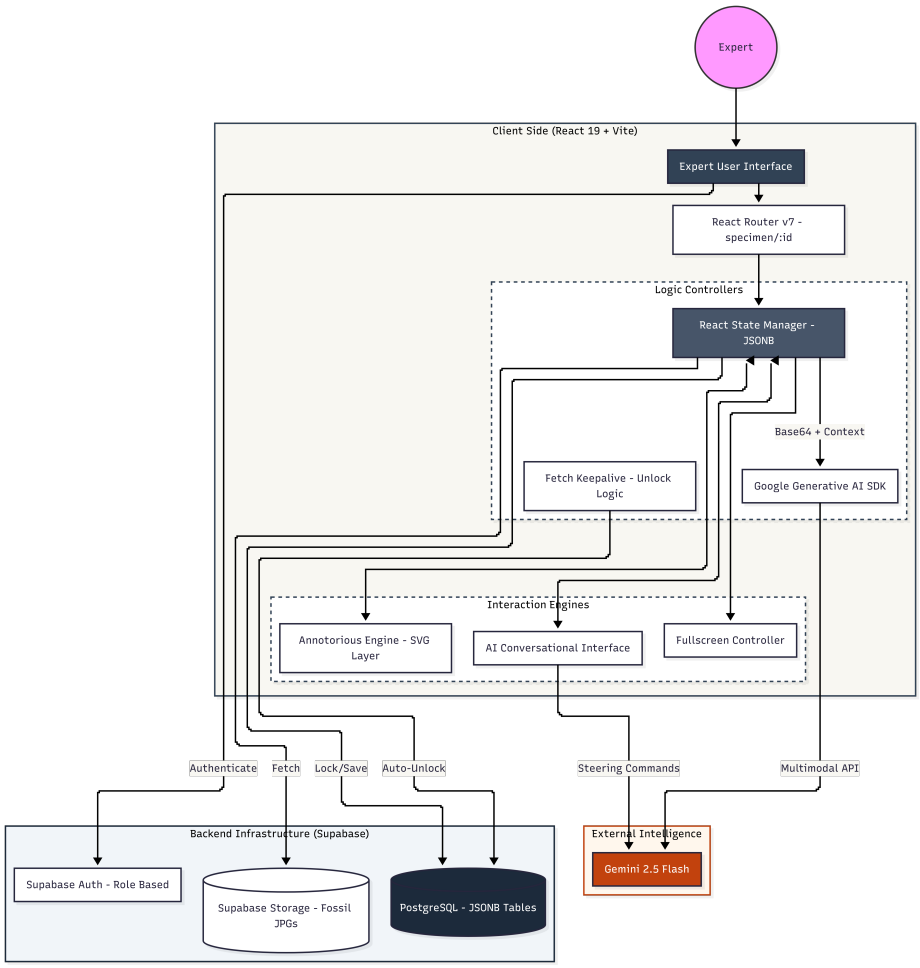


Fig. 16: **Technical architecture diagram of the annotation framework.**



Fig. 17: **Database schema of GoldenFossils.**



Fig. 18: **Length measurement error (mm) from the ground truth.**

Fig. 19: **Specimens exhibiting universal identification failure across all VLMs.** These samples represent the highest degree of morphological convergence and taphonomic degradation in the dataset.



(a) Mandible fragment (Key 92)



(b) Bone splinter



(c) Hollow diaphysis



(d) Distal astragalus



(e) Weathered phalanx



(f) Molar fragment

Note: All six specimens were consistently misidentified.