Star Light, Star Bright: Can Multimodal Models Parse an Astronomy Visual Question Answering Dataset Right?

J. P. Naiman
School of Information Sciences
University of Illinois, Urbana-Champaign, IL 61820, USA
jnaiman@illinois.edu

Abstract

Figures are essential for scientific communication, and the field of Visual Question Answering (VQA) aims to test the figure understanding capabilities of large multimodal models (LMMs). Here, we investigate the accuracy of several low-cost LMMs on domain-specific VQA tasks for astronomical *images of the sky*. As instrument for this task we build a domain-specific VQA dataset comprising 2001 partially-synthetic figures divided evenly between contour plots, real telescope images drawn from NASA's SkyView, and "fake" sky images which share the axes and angular scales of the real images but draw their colors from a gaussian mixture model, alongside 47357 questions. Each figure is released as pristine and with realistic page-aging degradations mimicking the historical astronomical corpus. Testing three low-cost proprietary LMMs, we find models frequently fail to answer domain-specific questions at high accuracy. An initial comparison with 104 human-verified, astrometrically tagged figures from historical journal publications shows similar median errors in angular extent. All code is available on GitHub: https://github.com/jnaiman/SkyImagesWASP2026 and dataset is available on HuggingFace: https://huggingface.co/datasets/ReadingTimeMachine/astro_sky_image_vqa.

1 Introduction

In academic articles, figures synthesize and interpret scientific information. Their centrality has made figure parsing a main task of pipelines that digitize scientific documents (e.g., document layout analysis, Binmakhashen and Mahmoud, 2019; Ding et al., 2026; Kumar and Lehal, 2025), alongside figure classification (Jobin et al., 2019) and dataset extraction (Davila et al., 2019). With the growing prevalence of large multimodal models (LMMs) in scientific practice, much recent work tests whether these models can parse figures and "understand" their meaning (Hoque et al., 2022; Farahani et al., 2023) – commonly through Visual Question Answering (VQA), in which a model is given an image and questions about the image. Initially focused on natural images (e.g., Kafle and Kanan, 2017), chart-based VQA is now an expanding field (e.g., Huang et al., 2024). Early studies covered simple bar, line and scatter charts (Kahou et al., 2018), while follow-up work added more complex chart types (e.g., funnel, Han et al., 2023) and plots whose relationship to the underlying data is indirect, such as histograms and contour plots (Naiman et al., 2026; Evans et al., 2026). Absent are domain-specific VQA tasks, which are necessary to probe realistic scientific LMM applications (Chang et al., 2022; Huang et al., 2024).

Astronomy relies on *images of the sky*, a figure type that would benefit from domain-specific VQA. Sky images share some complexity with contour plots, mapping light intensity in a wavelength range to a color rather than a spatial scale (Evans et al., 2026), but encode position in right-ascension (RA) and declination (DEC) – a non-Cartesian system that can deviate substantially from a rectangular grid over large sky areas. Historical sky figures (published before $\sim$1997) add a further complication: many are grayscale rather than color, and scanning for upload to digital archives introduces errors of its own (Grant et al., 2000; Baird, 2007).

To the author's knowledge, there are no VQA datasets centered on images of the sky, for either present-day or historical figures. However, benchmarks focused on astronomy have recently been developed for other AI and machine learning (ML) tasks. AstroVisBench (Joseph et al., 2025) evaluates whether models can write the code that produces astronomical visualizations, and AstroMMBench (Shi et al., 2025) evaluates domain-specific reasoning over curated astronomical imagery via multiple-choice questions. However, neither mea-

Proceedings of the 4th Workshop on Artificial Intelligence for Scientific Publications, pages 180–191
November 10, 2026 ©2026 Association for Computational Linguistics

sures LMM free-form answers requiring calculations with a known ground truth.

Our research question is: *can low-cost LMMs perform domain-specific parsing of astronomical sky figures – including distinguishing real and synthetic sky sources – and does this degrade for historical-style figures?* To answer it, we present a VQA dataset focused on images of the sky in both present-day and historical contexts, alongside two visually similar plot types: contour plots drawn from several two-dimensional distributions, and "fake" sky images that keep their RA/DEC axes but draw their colors from a gaussian mixture model (GMM) rather than from a telescope. The data is used to test low-cost proprietary LMMs, comparing accuracy between original ("pristine") figures and those modified by realistic aging processes. Code, images, question sets and LMM answers are released for use by the community.

2 Dataset Construction

2.1 Figure Generation

Single-panel figures are generated following the work of Naiman et al. (2026); Evans et al. (2026) and are rendered with matplotlib (Hunter, 2007). Style, colormap, DPI, aspect ratio, font and font size are randomly selected; titles and axis labels are drawn from a word pool mined from the arχiv, with inline mathematics substituted at a fixed rate (Naiman et al., 2023). Alongside each image the underlying data array, sampled generation parameters, and pixel-coordinate bounding boxes for the plot area, axes, tick labels, title, axis labels, colorbar and colorbar label are saved. Figures are rejected and re-sampled if labels run off the canvas, if bounding boxes overlap beyond a tolerance, or if the plot area falls below a fraction of the figure.

Figure types are sampled evenly between contour plots with colors from various distributions (random, linear, gaussian mixture model) (Evans et al., 2026), images of the sky taken from telescopes, and images of the sky created by a gaussian mixture model. The GMM-sky images serve as a control: real-sky images do not provide figures with realistic axes, angular scale, rendering and color-scale distribution without a telescope signal. Comparing answers between the two families isolates whether errors stem from the domain-specific RA/DEC axis conventions (shared by both) or from the image source (which differs).

For both types of images of the sky, axes are stored as world coordinate system (WCS) with RA/DEC tick labels in several styles and a sampled epoch (e.g., J2000, B1950). A (target, wavelength) pair draws from a set of objects from the astronomical literature (targets described in Naiman 2026), and retrieved from sky surveys within NASA's SkyView (McGlynn et al., 1996) through astroquery (Ginsburg et al., 2019), with WCS handling via astropy (Astropy Collaboration, 2022).

The physical size of an image centered on a target on the sky can change across surveys and applications. To mimic a realistic scale of a sky image shown in an article, when SIMBAD (Wenger et al., 2000) sky-size for images (angular field of view, FOV) is available for a target, we use this to set the range of RA/DEC from which to extract an image. When absent, we make use of the FOV and angular resolution information stored in SkyView for each sky survey and select RA/DEC ranges from 50 times the smallest resolution up to 0.75 times the FOV. GMM-generated sky images follow the same distribution of angular sizes as the real sky images. Real-sky cutouts are cached at 300×300 and square. To create a set of realistic aspect ratios (and to match with contour plot parameters), the grid is sampled (n_x from 50–300, $n_y = n_x/\text{aspect}$). The index map between initial and rescaled pixels is exactly affine, such that arcsec/pixel stays isotropic and the center is preserved. A logarithmic stretch with clipping at the 99th percentile and its inversion is applied to both GMM- and real-sky images – the net effect removes extreme outliers while leaving the bulk of the distribution on its original scale.

Units are not included in colorbars. Only $\sim$20% of retrieved real-sky images declare a form of physical units (e.g., BUNIT – Jy/beam, MJy/sr, K, Rayleighs), and typically several data processing steps are applied between raw sky image data and published real-sky figures (e.g., Starck and Murtagh, 2006). A fuller pipeline of image processing (e.g., Joseph et al., 2025) to more closely align with the domain science (and thus, physical colorbar units) is left as future work.

Our synthetically generated figures trade the visual realism of a human-collected and annotated figure dataset for the detail of access to underlying data and the precision of bounding box locations (Methani et al., 2020; Wang et al., 2024; Masry et al., 2025; Naiman et al., 2026). Our inclusion of telescope-derived images of the sky with realistic angular scales acts as a compromise between a fully-synthetic and a hand-annotated figure dataset.

2.1.1 Page Aging of Figures

The figure generation pipeline described previously produces pristine scientific charts; however, much of the astronomical literature is not "born digital", but instead consists of human-scanned copies of originally paper-only journal articles. Given that the storage and scanning procedures for historical images can degrade the uploaded images, we also build a copy of each figure with several realistic "page aging" image augmentations. We make use of the Augraphy (Groleau et al., 2023) package, which models damage as accumulating in three phases – *ink* (e.g., ink bleed), *paper* (e.g., color, texture) and *post* (e.g., stains, folding). Our application pulls randomly from this set of effects following Guan et al. (2025) to mimic a modified form of the "archive" set of effects from s1mb1o (2025), which is meant to align with expected degradations resulting from scanning historical documents. As many of the historical articles in the field of astronomy were scanned at high resolution by trained archivists (Grant et al., 2000), the uploaded documents retain a lower level of page augmentations – denoted as "archive-light" in this paper. Finally, we transform 2/3 of the images to grayscale to mimic the color palette of historical articles. Not all of the randomly selected original colormaps are suitable to transform to grayscale which adds a further source of error in the aged figures.

An example of a real-sky image (left panel) and aged counterpart (right panel) is shown in Figure 1.

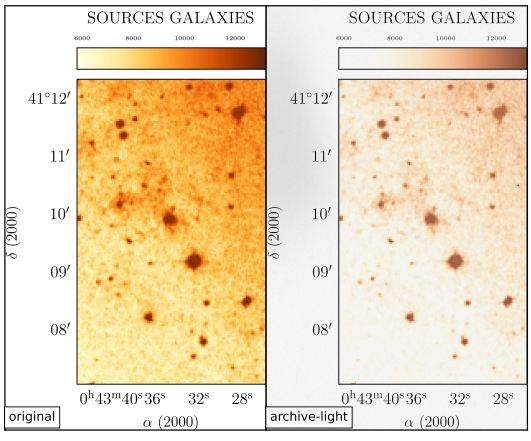


Figure 1: A dataset figure before (left) and after (right) archive-light aging.

2.2 Question Generation

Questions are generated for each figure using the four-level complexity model of Lundgard and Satyanarayan (2022), as adapted to chart-based VQA in other works (Tang et al., 2023; Naiman et al., 2026; Evans et al., 2026). We generate questions at Levels 1–3, spanning what is directly readable off the figure, what requires reading values, and what requires inference beyond a single statistic. Level 4 requires external context or domain expertise and is out of scope for this work. Example questions are listed in Table 1; Appendix B gives further detail and the accessibility motivation for this question scaffolding.

3 Results and Analysis

From our initial dataset of 2001 images, evenly spread between contours, real-sky images, and GMM-sky images and a total of 47357 questions, we randomly select 255 images (85 in each category) for evaluation with three proprietary LMMs, yielding 6035 questions per model. We use each in a low-cost configuration: GPT-5.4-nano (OpenAI, 2026) at default reasoning and verbosity (GPT), Gemini-3.5-flash-lite (Google DeepMind, 2026) with thinking_level = minimal (Gemini), and Claude-Haiku-4.5 (Anthropic, 2025) with thinking disabled (Claude). We restrict this preliminary run to the low-cost tier, where Naiman et al. (2026) found accuracy varied most. These are each provider's lower-cost reasoning setting and are not strictly equivalent across providers. The models selection and question scale assumes a cost-sensitive archiving institution, while holding the 95% Wilson interval (Kahouadji, 2025) on a single accuracy to at most ±10.4 percentage points (pp) with 85 figures per category, and ±7.4 pp for the 170 pooled sky images – enough to resolve the large differences discussed below, but not fine-grained model rankings. Questions are issued one per API call, so each answer is an isolated zero-shot observation. Each model is run over two conditions — the original figures and archive-light. Answers are requested as JSON with entries for raw answer and explanation. Only 48 of the 36210 answers (0.13%) fall outside this pair and are unparsable at scale – 21 that are not valid JSON and 27 explicit `null` refusals; per-model counts are given in Appendix A. Here, we report question types that probe sky-image features – plot type, provenance and angular scale in RA/DEC.

Plot type and Provenance. Figure 2 shows the accuracy of answers to "What is the plot type of the plot in this figure? Please choose the plot type from the following list: [contour, image of the sky]." and,

for sky images, "What is the underlying distribution used to create the data in this figure? Please choose the distribution from the following list: [gaussian mixture model, real image of the sky]."

These questions test two sources of possible confusion in the models – the presence of an image-based figure (contour) and the presence of the specific x/y labels of RA/DEC (GMM-sky). The tests are motivated by work showing scholarly documents can be a carrier for problematic adversarial content aimed at AI-assisted pipelines (Lin, 2026; Nagaraja et al., 2026). GMM-sky images are sampled to the same pixel, angular and color-scale distributions as the real ones, with parameters chosen to span the range from point sources to diffuse emission. We did not pursue statistical indistinguishability as the aim was to remove obvious artifacts while preserving the structural differences the task is meant to probe. A gradient-boosted classifier on low-level image features (e.g., field size) separates real from GMM-sky images with 85% accuracy (Appendix D). The two families are thus distinguishable from image content, so low LMM accuracy reflects model capability rather than an impossible task.

The accuracy shown in Figure 2 is measured as a simple binary (yes/no) for both original and archive-light historical figures. The left, plot type panels of Figure 2 show that when confronted with a contour plot Claude does worse than guessing (15% original, 22% archive-light) and GPT only modestly better (66% and 56%), except Gemini which shows ≈80% accuracy for both original and archive-light figures. Gemini and Claude recover sky-image plot types almost always, but this is to be expected from the over-guessing of this type of figure evident in the contour plot panels; GPT, by contrast, labels only 53% of original sky images correctly. For sky image source, the models clearly favor guessing "real image of the sky" and frequently miss when the image source is a GMM distribution. Only Claude returned unparsable answers here (Appendix A). Across models and conditions, GMM-sky images are identified correctly only 4–14% of the time, against 80–98% for real-sky images. With $n = 85$ per cell, differences of $\lesssim$10 pp lie within the 95% interval of a single cell, but the largest gaps do not: for example, Claude's 15% versus Gemini's 81% on contour plots gives $z = 8.6$ ($p < 10^{-16}$, simple two-proportion test).

Our dataset includes open-ended (no lists provided) versions of several questions. For example,

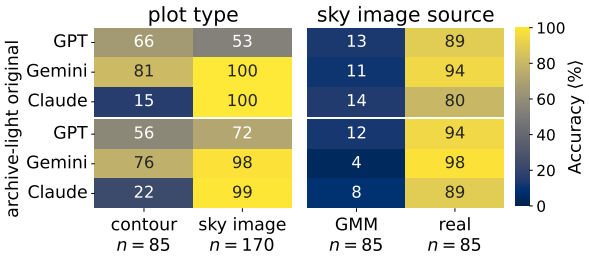


Figure 2: Accuracy of identifying plot type (left) and real vs generated sky (right) for the three low-cost models. Models are offered explicit choices for each.

in the plot type question raw accuracy is near zero for all three models as they answer with strings such as "astronomical survey image" that describe the figure correctly and never match the ground-truth label. Further parsing of open-ended answers is left for future work.

Angular height. Figure 3 shows the accuracy of answers to the question "What is the angular height of the sky region shown in this figure, measured along the declination axis? Give the value in arcminutes." This requires reading the y-axis of the figure and translating the domain-specific measure of angular size (see Figure 1) into another typical unit of angle in astronomy (arcmin, 1/60°).

The accuracy is measured by the log-ratio, $\mathcal{R} = \log_{10}(\hat{y}/y)$, defined for $y > 0$ and $\hat{y} > 0$ and scales the magnitude of a prediction similarly for under or over estimates. The metric cannot be used when LMM and ground truth (GT) answers have different signs which occurs in our dataset for a single GPT, archive-light answer.

Questions about angular scale probe how much detail is expected (angular resolution) and what other targets fall within the field (angular height/width). While Gemini does the best overall (peaked around $\mathcal{R} \approx 0$ in Figure 3), the other models have larger spreads tending toward overestimation with GPT having 53% (60%) of its responses at $> +1$ dex and Claude having 32% (32%) for pristine (arch-light/archive-light) figures. Initial parsing of explanations returned by the models suggests the slight overestimations ($0.5 < \mathcal{R} < 1$) stem from imprecise calculation of the y-axis ranges and rounding of answers. Larger errors ($\mathcal{R} > 1$) often result from misreading of the y-axis tick marks due to domain-specific norms (e.g., the bottom y-tick in Figure 1 is read as $08°00'$ instead of $08'$). Splitting Figure 3 by source shows the GMM-sky images are potentially easier for this axis-reading question: on

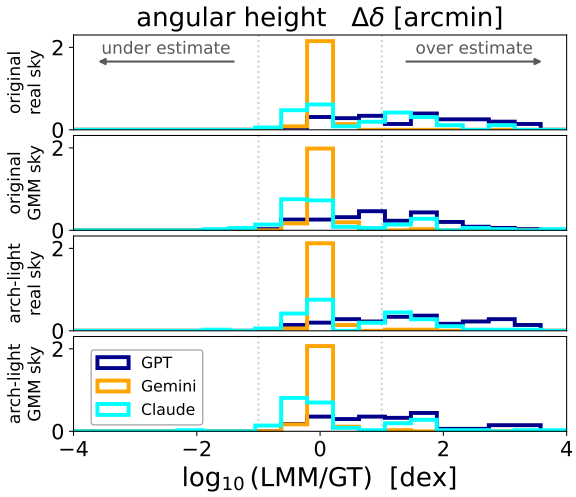


Figure 3: Error distribution of angular height for GMM- and real-sky images measured as log-ratio, $\mathcal{R}$. Zero indicates agreement with ground truth, ± 1 indicates predictions too large or too small by a factor of ten.

the original figures, Claude overestimates by more than a factor of ten significantly more often on real-sky than on GMM-sky figures (41% vs. 24%, $p = 0.01$), and GPT shows the same direction, though not significantly (59% vs. 47%, $p = 0.12$).

Published figures. To test whether results on our figures carry over to published ones, we collect sky images with human-verified astrometry from the AAS Astronomy Image Explorer (AIE) (American Astronomical Society and IOP Publishing, 2014; Naiman, 2020), which indexes figures from the Astrophysical and Astronomical Journals and includes Astronomy Visualization Metadata (AVM) in the WorldWide Telescope convention (Rosenfield et al., 2018). The AVM stores central RA/DEC and pixel scale, but does not include figure bounding box. Following an initial pass in OpenCV for potential image corners (Bradski, 2000; Suzuki and Abe, 1985; Douglas and Peucker, 1973) bounding boxes and figure canvas spans were verified by hand using Label Studio (Tkachenko et al., 2020). Along with automated checks of AVM to SIMBAD's position of the object named in its publication's title (if available) and checks for coordinate duplication or nonphysical field sizes, human verification reduced the AVM-tagged set of figures from 163 to 104.

As AVM tags exclude information for other figure elements (e.g., colorbar), questions are limited to angular scale and passed to our best-performing (Gemini) model. Figure 4 compares angular-width errors (change in RA de-projected from spherical coordinates, see Appendix C). Answers (103, 1 unparsable) are similar across synthetic (real-sky) and published figures: median $\mathcal{R} = -0.08$ dex against -0.05 dex for archive-light, with 55% vs. 51% of answers within a factor of two.

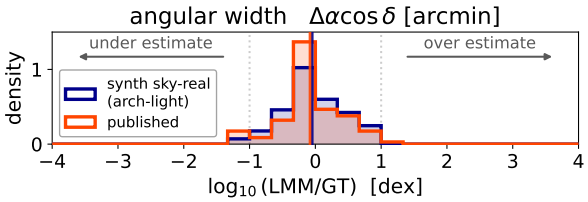


Figure 4: Log-ratio error, $\mathcal{R}$, of Gemini's angular-width answers for real-sky figures (archive-light, blue; $n = 85$) and for 103 human-verified published figures from the Astronomy Image Explorer (orange).

4 Conclusions and Future Work

We have introduced a domain-specific chart VQA dataset for astronomical figure types: contour plots and images of the sky. Sky panels are drawn both from real telescope observations and from GMMs, generated for equal numbers in contour, real- and GMM-sky image categories. We include both pristine images and "aged" images to mimic those in the historical astrophysics corpus.

Across three proprietary low-cost LMMs we find figure-type identification often fails to distinguish between sky and contour plots and, for sky images, between real and GMM images. Models tend to overestimate the angular heights of sky regions.

On a small set of human-verified published figures, Gemini's angular-width answers match its median behavior on our partially synthetic figures. We detect no difference in accuracy between pristine and aged images apart from one cell (GPT labels aged sky images correctly more often – 53% vs. 72%, $p < 0.001$). At $n = 85$ per model/category we can exclude only aging effects larger than $\sim$21 pp at 80% power. Similar questions asked with less domain-specific charts do not display such large inaccuracies (e.g., Naiman et al., 2026). These results motivate the need for the development of further domain-specific VQA datasets (Chang et al., 2022; Huang et al., 2024; Naiman et al., 2026).

Future work could focus on a more domain-accurate image processing pipeline, target other types of astronomy-specific figures (e.g., spectra), and investigate the explanations and open-ended answers stored as part of the LMM responses in our dataset which we did not explore here.

Limitations

The figures are generated, not collected. The arrays inside the real sky panels are genuine survey cutouts, but the figure drawn around them is not a published figure, and lacks conventions real ones carry — domain-standard colormaps, compass roses, beam ellipses, overlaid catalog positions. Our published-figure comparison suggests angular extent transfers in the median, but it covers only one model and two questions (angular width, and angular height in Appendix C). The 104 AVM-tagged figures are also a small, non-random sample of the sky images in the historical literature – those an archive happened to tag with astrometry (Naiman, 2020) – so the comparison should be read as an example of how results on our figures carry over, not as a definitive measure of LMM accuracy on published sky figures.

Matching the synthetic sky fields to the real ones removes the obvious cues to provenance, but does not make the two families indistinguishable and was not intended to. The provenance question we pose is therefore hard rather than impossible, and the accuracies we report should be read against a task that remains somewhat solvable in principle from image content alone (see Appendix D).

We evaluate only the lower-cost configuration of three proprietary models, on a single aging preset and these configurations are not strictly equivalent in reasoning effort across providers. Open-weight models, higher-reasoning configurations and astronomy-tuned models are untested, and model versions and prices change quickly enough that these particular numbers have a short shelf life.

Each condition is a single run over 255 of the 2001 figures. Per-cell uncertainties are up to ± 10.4 pp, so small differences between models, or between pristine and aged figures, should not be overinterpreted.

Our published-figure comparison is small (104 human-verified figures, one model) and limited to the geometric questions that astrometric metadata can ground; colormap, provenance and value questions have no ground truth for figures we did not draw. Its angular extents rely on human-verified plot boxes, and pixel scale cannot be compared at all, because the published astrometry counts scan pixels rather than data pixels. We do not include a human baseline on the real-versus-GMM question, which would require a hand-annotated corpus beyond the scope of this short paper.

The dataset is designed as an evaluation benchmark. We do not test the use of the synthetic images for training or data augmentation, though it is not outside the realm of possibility that this dataset could be used as training or pre-training for a variety of models.

Scoring is exact match after normalization for categorical answers and a log-ratio for numeric ones. As the open-ended plot-type results show in our preliminary investigation, exact match understates a model that is substantively right in unanticipated vocabulary. The questions themselves are generated from templates over known parameters: correct by construction, but not always a sample of what a working astronomer would ask.

Ethics Statement

The sky cutouts used here are public data retrieved from NASA's SkyView service and are redistributed within the dataset in derived form, with full provenance metadata. The published figures used for comparison may remain under their journals' copyright; we release only their metadata, questions and model answers, with a script that retrieves the images from the Astronomy Image Explorer. The synthetic sky panels in this dataset are visually plausible but are not observations, and are labeled as such in every record; we note this explicitly because the same matching procedure that makes them useful as a control would make unlabeled synthetic sky images misleading if separated from their metadata.

More broadly, image falsification is a recognized form of scientific misconduct[1] (a growing issue along with adversarial image and text content injection (Lin, 2026; Nagaraja et al., 2026)), and our synthetic sky images are, by design, plausible enough that the models we test usually label them as real. The procedure that makes them a useful control could be misused to fabricate observations in a manuscript, and our results suggest that low-cost LMMs would not reliably flag them. We mitigate this by labeling every synthetic image in its record and releasing the generation code and parameters so that any image can be traced, and we encourage use of the dataset to develop provenance detectors rather than as a source of figures, though we cannot limit the codebase's use toward generation of synthetic figures passed off as authentic.

[1] https://bps.stanford.edu/home/instances-scientific-misconduct/falsification-images

AI was used to port code from prior work to the current repository and help to provide documentation. AI was also used to write analysis and verification code (e.g., the classifier baseline, statistical checks, and table and cost generation, partially adapted from author-written code), all of which was reviewed by the author. It was additionally used to assist in the literature review (all sources verified by humans), with a draft of the methods section, appendices, and the abstract, and with the writing of the Limitations and Ethics Statement section. It was also used to help shorten the paper to fit page requirements and to help draft revisions responding to reviewer comments; all revised text was checked by the author. AI was trialed for other parts of the writing, but was not found useful in other sections so was abandoned for the main section of the paper.

References

American Astronomical Society and IOP Publishing. 2014. Astronomy image explorer. https://www.astroexplorer.org/. Launched 2014; accessed 2026-10-08.

Anthropic. 2025. Claude Haiku 4.5. https://www.anthropic.com/claude/haiku. Large language model. Released October 2025. API model identifier: `claude-haiku-4-5`.

Astropy Collaboration. 2022. The astropy project: Sustaining and growing a community-oriented open-source project and the latest major release (v5.0) of the core package. *The Astrophysical Journal*, 935(2):167.

Henry S. Baird. 2007. The state of the art of document image degradation modelling. In Bidyut B. Chaudhuri, editor, *Digital Document Processing: Major Directions and Recent Advances*, Advances in Pattern Recognition, pages 261–279. Springer, London.

Galal M. Binmakhashen and Sabri A. Mahmoud. 2019. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6).

Gary Bradski. 2000. The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. 2022. Mapqa: A dataset for question answering on choropleth maps. In *NeurIPS 2022 First Table Representation Workshop*.

Kenny Davila, Bhargava Urala Kota, Srirangaraj Setlur, Venu Govindaraju, Christopher Tensmeyer, Sumit Shekhar, and Ritwick Chaudhry. 2019. ICDAR 2019 competition on harvesting raw tables from infographics (CHART-Infographics). In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1594–1599, Sydney, Australia. IEEE.

Yihao Ding, Soyeon Caren Han, Jean Lee, and Eduard Hovy. 2026. Deep learning based visually rich document content understanding: a survey. *Artificial Intelligence Review*, 59(4):114. Open access.

David H. Douglas and Thomas K. Peucker. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122.

Daniel John Evans, Jill P. Naiman, JooYoung Seo, and Stephen J. Downie. 2026. Do you "see" what I "see"? A Multi-panel Visual Question Answer Dataset for Large Language Model Chart Analysis. In *The 30th International Conference on Theory and Practice of Digital Libraries*, The 30th International Conference on Theory and Practice of Digital Libraries.

Ali Mazraeh Farahani, Peyman Adibi, Mohammad Saeed Ehsani, Hans-Peter Hutter, and Alireza Darvishy. 2023. Automatic chart understanding: A review. *IEEE Access*, 11:76202–76221.

Adam Ginsburg, Brigitta M. Sipőcz, C. E. Brasseur, Philip S. Cowperthwaite, Matthew W. Craig, Christoph Deil, James Guillochon, Giannina Guzman, Simon Liedtke, Pey Lian Lim, Kelly E. Lockhart, Michael Mommert, Brett M. Morris, Henrik Norman, Madhura Parikh, Magnus V. Persson, Thomas P. Robitaille, Juan-Carlos Segovia, Leo P. Singer, and 4 others. 2019. astroquery: An astronomical web-querying package in python. *The Astronomical Journal*, 157(3):98.

Google DeepMind. 2026. Gemini 3.1 Pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Large language model. Released February 19, 2026. API model identifier: `gemini-3.1-pro`.

Carolyn S. Grant, Alberto Accomazzi, Guenther Eichhorn, Michael J. Kurtz, and Stephen S. Murray. 2000. The NASA Astrophysics Data System: Data holdings. *Astronomy and Astrophysics Supplement Series*, 143(1):111–135.

Alexander Groleau, Kok Wei Chee, Stefan Larson, Samay Maini, and Jonathan Boarman. 2023. Augraphy: A data augmentation library for document images. In *Document Analysis and Recognition – ICDAR 2023*, volume 14191 of *Lecture Notes in Computer Science*, pages 384–401. Springer Nature Switzerland.

Shuhao Guan, Moule Lin, Cheng Xu, Xinyi Liu, Jinman Zhao, Jiexin Fan, Qi Xu, and Derek Greene. 2025. PreP-OCR: A complete pipeline for document image restoration and enhanced OCR accuracy. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*.

Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. 2022. Chart question answering: State of the art and future directions. *Computer Graphics Forum*, 41(3):555–572. EuroVis 2022 State-of-the-Art Report.

Kung-Hsiang Huang, Hou Pong Chan, Yi R. Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models. *arXiv preprint*. ArXiv:2403.12027 [cs] version: 2.

John D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

K. V. Jobin, Ajoy Mondal, and C. V. Jawahar. 2019. DocFigure: A dataset for scientific document figure classification. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 1, pages 74–79. IEEE.

Sebastian Antony Joseph, Syed Murtaza Husain, Stella S. R. Offner, Stéphanie Juneau, Paul Torrey, Adam S. Bolton, Juan P. Farias, Niall Gaffney, Greg Durrett, and Junyi Jessy Li. 2025. AstroVisBench: A code benchmark for scientific computing and visualization in astronomy. *Preprint*, arXiv:2505.20538. NeurIPS 2025 Datasets and Benchmarks Track.

Kushal Kafle and Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20. Language in Vision.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning. *Preprint*, arXiv:1710.07300.

Nabil Kahouadji. 2025. A comprehensive comparison of the wald, wilson, and adjusted wilson confidence intervals for proportions. *arXiv preprint arXiv:2508.10223*.

Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim. 2023. Exploring Chart Question Answering for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, pages 1–15, New York, NY, USA. Association for Computing Machinery.

Atul Kumar and Gurpreet Singh Lehal. 2025. Exploring document layout analysis: A review. In *Innovative Computing and Communications: Proceedings of ICICC 2025*, volume 1438 of *Lecture Notes in Networks and Systems*, pages 367–383, Singapore. Springer.

Zhicheng Lin. 2026. Hidden prompts in manuscripts exploit AI-assisted peer review. *Communications of the ACM*, 69(7):53–56.

Alan Lundgard and Arvind Satyanarayan. 2022. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Transactions on Visualization & Computer Graphics (Proc. IEEE VIS)*.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. 2025. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. *arXiv preprint*. ArXiv:2504.05506 [cs] version: 2.

T. McGlynn, K. Scollick, and N. White. 1996. SkyView: The multi-wavelength sky on the internet. In *New Horizons from Multi-Wavelength Sky Surveys*, number 179 in IAU Symposium, page 465. Kluwer Academic Publishers.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over Scientific Plots. pages 1527–1536.

Neha Nagaraja, Lan Zhang, Zhilong Wang, Bo Zhang, and Pawan Patil. 2026. Image-based prompt injection: Hijacking multimodal LLMs through visually embedded adversarial instructions. In *3rd International Conference on Foundation and Large Language Models (FLLM)*, Vienna, Austria.

J. P. Naiman, Morgan G. Cosillo, Peter K. G. Williams, and Alyssa Goodman. 2023. Large synthetic data from the arχiv for ocr post correction of historic scientific articles. In *Linking Theory and Practice of Digital Libraries: 27th International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, September 26–29, 2023, Proceedings*, page 265–274, Berlin, Heidelberg. Springer-Verlag.

Jill Naiman. 2020. The Reading Time Machine: Transforming Astrophysical Literature into Actionable Data. NASA Proposal id.20-ADAP20-225.

Jill P. Naiman. 2026. Astronomical targets from the literature for sky-image figure generation. Submitted to ACM Transactions on AI for Science.

Jill P. Naiman, Daniel J. Evans, and JooYoung Seo. 2026. What Lies Beneath: A Call for Distribution-based Visual Question & Answer Datasets. *arXiv preprint*. ArXiv:2601.22218 [cs].

OpenAI. 2026. ChatGPT (GPT-5.4). `https://openai.com/index/introducing-gpt-5-4/`. Large language model. Released March 5, 2026. API model identifier: `gpt-5.4`.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Philip Rosenfield, Jonathan Fay, Ronald K. Gilchrist, Chenzhou Cui, A. David Weigel, Thomas Robitaille, Oderah Justin Otor, Alyssa Goodman, et al. 2018. AAS WorldWide Telescope: A seamless, cross-platform data visualization engine for astronomy research, education, and democratizing data. *The Astrophysical Journal Supplement Series*, 236(1):22.

s1mb1o. 2025. scan-simulator: A synthetic degradation pipeline for scans, photos and aged documents. `https://github.com/s1mb1o/scan-simulator`. Software, MIT license.

JooYoung Seo, Yilin Xia, Bongshin Lee, Sean McCurry, and Yu Jun Yam. 2024. MAIDR: Making Statistical Visualizations Accessible with Multimodal Data Representation. ArXiv:2403.00717 [cs].

Jinghang Shi, Xiaoyu Tang, Yang Huang, Yuyang Li, Xiao Kong, Yanxia Zhang, and Caizhan Yue. 2025. AstroMMBench: A benchmark for evaluating multimodal large language models capabilities in astronomy. *Preprint*, arXiv:2510.00063.

Jean-Luc Starck and Fionn Murtagh. 2006. *Astronomical Image and Data Analysis*, 2 edition. Astronomy and Astrophysics Library. Springer, Berlin, Heidelberg.

Satoshi Suzuki and Keiichi Abe. 1985. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A Benchmark for Semantically Rich Chart Captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. `https://github.com/HumanSignal/label-studio`. Open source software, version 1.23.2 used.

Lucy Lu Wang, Isabel Cachola, Jonathan Bragg, Evie Yu-Yen Cheng, Chelsea Haupt, Matt Latzke, Bailey Kuehl, Madeleine van Zuylen, Linda Wagner, and Daniel S. Weld. 2021. Improving the accessibility of scientific documents: Current state, user needs, and a system solution to enhance scientific PDF accessibility for blind and low vision users. *arXiv preprint arXiv:2105.00076*.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:113569–113697.

M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, and R. Monier. 2000. The SIMBAD astronomical database. The CDS reference database for astronomical objects. *Astron. Astrophys. Suppl. Ser.*, 143:9–22.

A Answer Parsing

Answers are requested as JSON with entries for the raw answer and an explanation. A small number fall outside that pair and cannot be parsed at scale. Out of 6035 questions per model, this occurred for 12 (22), 4 (3) and 5 (2) answers for GPT, Gemini and Claude respectively, where the first number is for the original images and the bracketed number for archive-light. For Gemini's and Claude's are responses that are not valid JSON: most break on the quote characters of sexagesimal tick labels (e.g., `12'00"`), and four of Claude's answer the provenance question in free text with no JSON at all. 27 of GPT's 34 are valid JSON with an explicit `null` answer, typically for color statistics it judges unrecoverable from the image. For the plot-type and provenance questions of Figure 2 specifically, only Claude returned unparsable answers: 3 for the original images and 1 for archive-light, out of 425 each.

B Question Levels

Questions follow the four-level figure description model of Lundgard and Satyanarayan (2022), adapted to chart-based VQA in Tang et al. (2023); Naiman et al. (2026); Evans et al. (2026). Level 1 covers what is directly readable off the figure (e.g., plot type); Level 2 requires reading values off the axes or colorbar (e.g., the median of the data along an axis, or the angular height of the field); Level 3 requires inference not reducible to a single statistic, such as which distribution family the underlying data came from or multi-step, complex, calculations. Level 4 requires external context or domain expertise and is not generated here.

Beyond quantifying question complexity, this scaffold matters for accessibility: many scientific figures remain unusable for blind and low-vision readers (Wang et al., 2021), and Levels 2 and

3 are the tiers those readers report as most useful (Lundgard and Satyanarayan, 2022; Seo et al., 2024; Kim et al., 2023).

Table 1 gives example questions for our figures; Table 2 lists every question asked of the published figures and where its ground truth comes from.

C Angular Extent of a Sky Figure

Positions on the sky are given in equatorial coordinates: right ascension α, measured eastward along the celestial equator, and declination δ, the angle north or south of it. By convention α is written in hours, minutes and seconds of time, with $24^{\mathrm{h}} = 360°$, so that $1^{\mathrm{h}} = 15°$, $1^{\mathrm{m}} = 15'$ and $1^{\mathrm{s}} = 15''$, while δ is written in degrees, arcminutes and arcseconds. There are two potential issues when reading a sky figure. First, a minute of right ascension is not an arcminute: an axis spanning 2^{m} of α covers $30'$. Second, lines of constant α converge toward the celestial poles, so the true angle subtended by an interval $\Delta\alpha$ shrinks with declination; along a line of constant δ it is $\Delta\alpha \cos\delta$. At $\delta = 60°$, one hour of right ascension spans $7.5°$ on the sky rather than $15°$. Angular height, the extent along the declination axis, needs neither correction, since $\Delta\delta$ is already an angle.

Our ground truth applies the $\cos\delta$ factor at the mid-declination of the displayed field, $w = (\alpha_{\max} - \alpha_{\min}) \cos\bar{\delta}$, whereas for the published figures width is the great-circle length of the plot box's top and bottom edges, measured through the AVM astrometry. The two definitions agree to within 2% for 94 of the 104 published figures (median 0.4%) and differ appreciably only for fields spanning several degrees, by up to 23% for a $16°$-wide field at $\delta \approx 58°$.

Angular height on published figures. Gemini was also asked the angular height of each published figure, worded identically to our figures (Table 2). Figure 5 compares the errors. In the median the two sets agree, as for width: $\mathcal{R} = -0.04$ dex for published figures against -0.05 dex for archive-light. Published figures scatter more, however: 76% of answers fall within a factor of two, against 94% for our figures ($p < 0.001$, two-proportion test), and 10% are too large by more than a factor of ten, against 2%. Since height needs no $\cos\delta$ correction, this extra scatter cannot come from the right ascension conventions above.

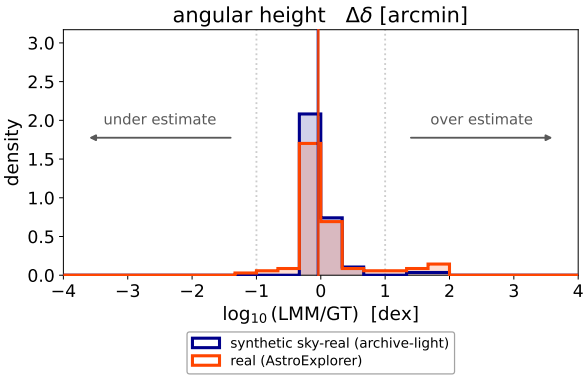


Figure 5: As Figure 4, for angular height: log-ratio error, $\mathcal{R}$, of Gemini's answers for our real-sky figures (archive-light, blue; $n = 85$) and for the 104 human-verified published figures (orange).

D Classifier Baseline for Provenance

To determine if the provenance question can be answered from image content, we train a gradient-boosted classifier (Pedregosa et al., 2011) to separate real-sky from GMM-sky figures using three low-level features of each figure's data array: the field size (the geometric mean of its angular width and height), the concentration of its emission (the share of the background-subtracted signal held by the brightest 5% of pixels), and the sign of the data minimum. Over all 1334 sky figures (667 of each family), 5-fold cross-validation gives 85% accuracy, against 50% for always guessing one family. Most of this comes from concentration, which alone reaches 80%: point-source-dominated real fields are more concentrated than the GMM fields. Field size and the sign of the minimum carry little information on their own, as the matching of the two families intends.

We leave out a fourth feature, the color span (the dynamic range of the data). GMM fields borrow their color ranges from real cutouts, and in this release the same ranges recur across many GMM figures: 667 GMM figures share 56 distinct spans, while real figures have 473. A classifier can exploit that regularity across the dataset – color span alone reaches 81% – but no single figure reveals it, so it says nothing about whether the provenance question is answerable from one image. The code is released with the dataset generation code with updates to pull a color range distribution more closely resembling that of the real-sky images intended in future releases.

Lvl	Question [choices offered]	Applies to	Example Answer
1	What is the plot type of the plot in this figure? [contour, image of the sky]	all plots	image of the sky
1	What is the colormap that was used in this figure?	all plots	rainbow_r
1	What coordinate epoch is given on the axis labels of this figure? Answer "none" if no epoch is stated.	GMM-sky, real-sky	J2000
1	What is the smallest unit of angle shown on the right ascension tick labels of this figure? [hours, minutes, seconds]	GMM-sky, real-sky	minutes
1	What is the minimum right ascension value covered by the right ascension axis of this figure – that is, the lower limit of the axis range, not of the data? Give the value in hours, minutes and seconds [...]	GMM-sky, real-sky	18h46m50.98s
1	What is the style of the contour plot? [image, contour lines, both]	contour	contour lines
2	What is the angular height of the sky region shown in this figure, measured along the declination axis? Give the value in arcminutes.	GMM-sky, real-sky, published	1.58
2	What is the angular width of the sky region shown in this figure, measured along the right ascension axis, as a true angle on the sky (i.e. including the cos(declination) factor)? Give the value in arcminutes.	GMM-sky, real-sky, published	1.88
2	What is the mean value of the data along the color-axis in this figure?	all plots	108.4
3	What is the underlying distribution used to create the data in this figure? [gaussian mixture model, real image of the sky]	GMM-sky, real-sky	real image of the sky
3	What is the underlying distribution used to create the data in this figure along the color-axis? [random, linear, gaussian mixture model]	contour	linear
3	Approximately what angular size does a single pixel of this sky image span? Give the value in arcseconds.	GMM-sky, real-sky	188.0
3	What is the right ascension of the centre of this sky image, in the J2000 (FK5) frame? [...] Give the value in decimal degrees.	published	91.95

Table 1: Example questions for each level (Lvl column), the plot types they are asked of, and an example ground-truth answer. Choices offered are given in brackets; long prompts are abridged ([...]). Plot types: contour, GMM-sky and real-sky are our figures; published are the human-verified Astronomy Image Explorer figures. Example answers come from single figures of each type (vqa_000001, GMM-sky; vqa_000002 and vqa_000004, real-sky; vqa_000007, contour; rsky_000001, published); numeric answers are in the units the question asks for.

E Archive Aging Details

Table 3 lists the archive-light preset implemented: 11 effects, their Augraphy phase, the independent probability with which each fires for a given figure, and the realized firing rate over the 255-figure corpus. Effects marked *soft* run with parameters pulled back from the Augraphy defaults — ink bleed at intensity 0.1–0.3, ink mottling at alpha 0.05–0.12, paper hue restricted to 28–40 at low saturation, stain blending at alpha 0.18, shadows at opacity 0.05–0.2, rotations restricted to $\pm 1^\circ$ rather than $\pm 3^\circ$, and JPEG quality held at 70–95. Folding and SubtleNoise run at their defaults but fire less often.

Relative to the full archive preset of s1mb1o (2025), three effects are dropped entirely — BleedThrough, LightingGradient and PageBorder. A further four effects available in Augraphy are excluded from both archive (s1mb1o, 2025) and our archive-light presets: Scribbles and WaterMark add marks unrelated to the figure, BookBinding bends the page until the plot is no longer readable as a plot, and Letterpress restyles the ink against the colormap. None of these are likely to apply to our figures as they do not typically mimic damage expected in archive scans (Grant et al., 2000).

For each figure we record the sampled plan — the effects that fired, their phase assignment, the grayscale flag, the page background color and the seed — in a manifest released with the dataset, so any aged image can be regenerated without redistributing it. The plan is recorded rather than recomputed at read time, so a manifest stays valid for the images it describes even if the preset is later changed.

F Cost Analysis

Table 4 gives the cost of running each model once over the evaluation, i.e. every question asked a single time per condition. For Gemini and Claude it is computed exactly from the token counts recorded in each API response; one question costs about $0.0007 with Gemini and $0.0025 with Claude. Our code does not recorde GPT's token usage, and its reasoning tokens are billed but not returned, so its cost is scaled from the provider's billed cost for its final runs: $3.34 for the 4970 calls covering

Lvl	Question	Ground truth	Range	Shared
2	What is the angular width of the sky region shown in this figure, measured along the right ascension axis, as a true angle on the sky (i.e. including the cos(declination) factor)? Give the value in arcminutes.	mean length of the plot box's top and bottom edges on the sky	0.33–957′	yes
2	What is the angular height of the sky region shown in this figure, measured along the declination axis? Give the value in arcminutes.	mean length of the plot box's left and right edges on the sky	0.33–1429′	yes
3	What is the right ascension of the centre of this sky image, in the J2000 (FK5) frame? [...] Give the value in decimal degrees.	sky position of the plot-box centre	23–346°	no
3	What is the declination of the centre of this sky image, in the J2000 (FK5) frame? [...] Give the value in decimal degrees.	sky position of the plot-box centre	−67–75°	no
3	Approximately what angular size does a single pixel of this sky image span? Give the value in arcseconds.	not released: scan pixels, not data pixels	–	yes

Table 2: Questions asked of the 104 human-verified published figures (Gemini only). Ground truth is measured through each figure's AVM astrometry over its human-verified plot box (Appendix C); the range is over all 104 figures, in arcminutes (′) or degrees. "Shared" marks questions worded identically to those asked of our figures, so their scores are directly comparable. Questions about other figure elements (colormap, colorbar statistics, provenance) have no ground truth in the metadata and are not asked; pixel scale was asked but is not scored.

Phase	Effect	p	fired
ink	InkBleed *soft*	0.50	55%
ink	InkMottling *soft*	0.30	34%
paper	ColorPaper *soft*	0.80	80%
paper	NoiseTexturize *soft*	0.50	55%
paper	BrightnessTexturize *soft*	0.50	49%
post	Stains *soft*	0.40	38%
post	Folding	0.25	25%
post	ShadowCast *soft*	0.25	24%
post	Geometric *soft*	0.50	49%
post	SubtleNoise	0.50	48%
post	Jpeg *soft*	0.60	59%

Table 3: The archive-light preset: firing probability and the rate realized over 255 figures. Each aged figure is also converted to grayscale with probability 2/3 (realized: 65%).

the last 105 figures of in each condition, or about $0.0007 per question. All costs use non-batch API pricing so larger datasets would likely be less expensive.

Model	Condition	In (M)	Out (M)	Cost ($)
GPT	both (est.)	–	–	8.11
Gemini	original	7.54	0.86	4.42
Gemini	archive-light	7.54	0.77	4.18
Claude	original	10.34	0.96	15.14
Claude	archive-light	10.35	0.95	15.11
Gemini	published	0.66	0.20	0.69

Table 4: Cost of one run of each model: tokens in and out (millions) and cost in USD at list prices per million input / output tokens (GPT $0.20 / $1.25; Gemini $0.30 / $2.50; Claude $1.00 / $5.00), checked 12 October 2026. Each condition is 6035 calls per model; the published-figure run is 520 calls (Gemini only). GPT, whose responses do not record token usage, is scaled from its provider's billed cost for the last 4970 calls (both conditions).