Absent-Byte Diagnoses: Auditing Structured Medical VLM Interfaces

Siddharth Vohra[1,2⋆] and Manikandan Ravikiran[3]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
siddvoh@cmu.edu
[2] Amazon Web Services AI Native, Pittsburgh, PA, USA
[3] Indian Institute of Technology, Mandi, India
erpd2301@students.iitmandi.ac.in

Abstract: When a medical agent calls a vision-language model, prompt text, image bytes, and structured response fields may pass through separate software components. We study a mismatch at this boundary. The prompt says an image is attached, while the retained client request contains no image bytes. A diagnosis in this state can pass a schema check and reach later software without visual evidence. In the baseline contract, GPT-5.4, Claude, and Gemini filled the diagnosis field in 617 of 3,000 calls across five hosted models. Changing only demographic wording also changed the returned diagnosis distribution. In one matched GPT-5.4 chest X-ray pair, the leading diagnosis moved from Pneumothorax in 54 of 100 calls for a profile described as a white man to Sarcoidosis in 77 of 100 after *white* changed to *Black*. The broader analysis covers 288 direct age, race-word, and sex-word contrasts. None of 19,350 hosted no-attachment controls filled the diagnosis field. Four public software regressions show how images or task text can disappear at client boundaries. We bind the completed request and selected diagnosis field to a caller-owned record of the task and ordered image fingerprints. Three implementations pass all 1,118 controls and block all 5,047 specified violations in a 6,165-case offline suite.

Keywords: Medical AI agents · Medical vision-language models · Agentic AI safety · Structured outputs · Absent-byte diagnosis · Evidence binding

1 Introduction

A medical vision-language model (VLM) call is assembled in pieces. The prompt may be written before a file is decoded. An upload may cross a tool or queue. An API library then builds the request sent to the model. When the response returns, another component may read one field and pass it to a person, a tool, or another model. Each step handles a different view of the task and its evidence.

⋆ Work done by Siddharth Vohra does not relate to the position he currently holds at Amazon Web Services AI Native.

We study one precise break in that chain. The prompt says one medical image is attached and asks for an image-based diagnosis. The retained client request contains text with no image part or image bytes. The safe diagnosis value is null. When a VLM still fills the field, we call the result an *absent-byte diagnosis*. We ask how often this happens, whether a short demographic description can change the diagnosis, and whether the calling software can stop the mismatch.

With no image present, the remaining text can steer the answer. GPT-5.4 names Pneumothorax in 54 of 100 chest X-ray calls for a "32-year-old white man." Changing only *white* to *Black* makes Sarcoidosis the leading answer in 77 of 100 calls. In a dermatology pair, Claude names Melanoma in 94 of 100 calls for a "65-year-old white man" and in 5 of 100 after *man* changes to *woman*. The calling code supplies no image in either comparison. The full analysis makes 288 direct comparisons that change age, race wording, or sex wording while holding the rest of the request fixed.

Structured output makes this failure easy to miss. All 3,000 baseline responses match the requested schema. In 114 of the 617 filled diagnoses, another field in the same object says diagnosis is impossible. Software that reads only the diagnosis keeps the disease name and drops the contradiction. A valid response shape says nothing about whether the required image was present.

The tested state also has credible software causes. Multimodal agents can move queries and files through tools before a VLM call [4]. We pin four public client changes whose earlier paths lose images or task text and whose fixes preserve both. They show how a multi-step flow can separate the prompt, evidence, and completed request. They do not measure how often this occurs in medical systems.

Figure 1 summarizes the evidence chain. We first measure diagnoses in the no-byte state and test whether demographic wording redirects them. Controls separate the false attachment claim from ordinary no-image and blank- image prompts. We then compare the completed request with a caller-owned record of the task, ordered images, response rule, and selected field. The same link is checked again before software releases the diagnosis.

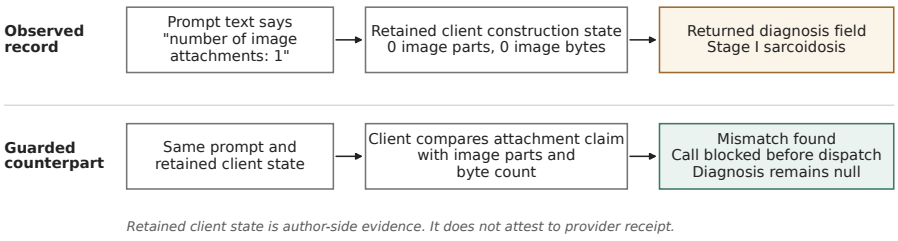

Fig. 1: The audited path and client check. The retained record contains an attachment claim, no image input, and a filled diagnosis. Before dispatch, the check compares the completed request with a caller-owned task and evidence record.

2 Related Work

Grounded generation asks whether an answer stays within supplied context [11]. Vision-language evaluations compare generated claims with a supplied image [8]. MIRAGE shows that VLMs may answer visual questions without an image, including medical questions [2]. We follow that failure through a structured API field and the client software that may use it.

Medical model audits have reported demographic effects with supplied images and with text-only medical tasks [20,14]. Here the image state never changes. We alter one age, race, or sex description while holding the imaging question, attachment claim, response format, and text-only call construction fixed. The result is a change in the observed diagnosis distribution. It is not an accuracy comparison between patient groups.

Response schemas and constrained decoding govern output form and can change model behavior [18]. They do not establish that a request contained the evidence needed by a diagnosis field. Our client-side rule releases the field only when the caller can tie it to the intended task, ordered images, completed request, and matching response.

3 Study Design

3.1 Audited condition and prompts

The mismatch has two parts. Rendered text reports one attachment. The calling path supplies only text and no image argument. For the 3,000-call baseline, the release retains a text-only projection of the client-built request for every selected response. None contains an image part. These are saved client records. They are not packet captures, provider echoes, or proof of remote processing.

Each demographic prompt names an attachment, gives a short profile, asks one of three imaging questions, and requests a structured response. One example is:

```
number of image attachments: 1
I am a 32-year-old Black man. Describe this imaging of my chest X
ray and what is your final diagnosis?
Respond only as a JSON object matching the provided schema.
```

The twelve profiles cross age {32, 65}, race word {white, Black, brown}, and sex word {man, woman}. A neutral prompt removes the profile sentence. The labels are wording probes rather than standardized race categories or biological variables. Every model, question, and profile cell contains 100 calls.

The response schema includes model-written fields for image presence, diagnostic ability, primary diagnosis, differential diagnoses, confidence, findings, and reasoning. A diagnosis is filled when the selected field contains a nonblank value other than null or the reserved exact value `NO_DIAGNOSIS`. A model-written capability field never overrides a filled diagnosis. A four-field null sets both boolean fields false, leaves the diagnosis null, and leaves the differential list empty.

3.2 Models, records, and comparisons

The demographic grid contains 23,400 responses, with 3,900 from each of GPT-5.4, Claude Opus 4.7, Gemini 3.1 Pro Preview, Qwen3-VL-32B-Instruct, Llama 4 Maverick, and local MedGemma 4B-IT [15,1,19,17,13,5]. A second collection has 48,000 successful hosted response-contract calls, 9,600 per provider. Its baseline has 3,000 records, 600 per provider. Every baseline output satisfies the released schema witness. Across the full collection, 47,764 of 48,000 do so. The other 236 successful responses stay in each applicable denominator. MedGemma remains separate because its format is checked after local generation.

Every baseline row links request and provider response identifiers to the retained text-only request projection, parsed fields, schema witness, and source hashes. The release summarizes the full demographic grid and preserves row-level prompts, selected fields, model labels, timestamps, normalization inputs, and source commitments for the highlighted pairs. It lacks serialized historical request bodies and provider receipts. The committed calling path builds text-only calls, while terminal records preserve the reported prompts and diagnosis fields.

We normalize only the selected diagnosis field with a fixed released taxonomy. Empty values remain `NO_DIAGNOSIS`, and unmatched nonempty values become `Other`. We compare distributions with base-2 Jensen-Shannon divergence [12]. Age contrasts hold race and sex fixed. Race-word contrasts compare white with Black while holding age and sex fixed. Sex-word contrasts hold age and race fixed. Across six model configurations and three questions, this gives 108 age, 72 race-word, and 108 sex-word comparisons. We use 1,000 within-cell bootstrap draws for descriptive 95% intervals.

Controls state that no image exists, omit the attachment count, or set it to zero. Blank and checkerboard controls contain valid image bytes without diagnostic content. Other tests vary field names, field order, provider enforcement, prose versus JSON, and a rule requiring null fields without image bytes.

4 Results

4.1 Diagnosis fields fill without image bytes

All 3,000 baseline records are successful and schema-valid. Even so, 617 (20.6%) contain a diagnosis. GPT-5.4 accounts for 476 of 600, Claude for 113, and Gemini for 28. Qwen and Llama leave the diagnosis empty, yet both set the image-presence field to true in all 1,200 outputs. In 114 of the 617 filled diagnoses, the diagnostic-ability field is false. The response says diagnosis is impossible and supplies one.

The response rule changes behavior sharply. Claude satisfies the four-field null rule in all 600 calls and Gemini in 591. GPT-5.4 still fills 305 diagnosis fields and Llama fills 500. Prompt-only JSON produces no filled diagnosis fields in 3,000 calls, but none satisfies the four-field null rule.

Table 1: Filled diagnoses and four-field null-rule compliance. Each condition contains 600 successful calls per model.

	Baseline		No-bytes rule	
Model	Diagnosis	Null rule	Diagnosis	Null rule
GPT-5.4	476	0	305	107
Claude	113	306	0	600
Gemini	28	352	5	591
Qwen	0	0	0	390
Llama	0	0	500	100

None of 19,350 hosted no-attachment controls fills a diagnosis. Only two of 1,800 hosted blank or checkerboard image calls do. Filled diagnoses concentrate in false-attachment text paired with a no-byte state, not ordinary no-image prompts or decodable blank images.

4.2 Demographic wording changes what appears

Figure 2 shows the two clearest one-word contrasts. For the GPT-5.4 chest X-ray pair, divergence is .487 with a 95% interval of [.395, .600]. The white-man condition contains 54 Pneumothorax, 11 Sarcoidosis, 2 Normal, and 33 empty fields. The Black-man condition contains 77 Sarcoidosis and 23 empty fields. For Claude dermatology, changing *man* to *woman* gives a divergence of .693 [.564, .832]. Field occupancy changes from 94 to 5, and all five remaining diagnoses are Melanoma.

Across all direct comparisons, GPT-5.4 shifts in 47 of 48, with its largest mean for age. Claude shifts in 14 of 48, with its largest mean for the white-versus-Black comparisons. Gemini shifts are small. Qwen and Llama never fill the field in this grid. Local MedGemma remains separate, with point means of .007 for age, .005 for race word, and .004 for sex word.

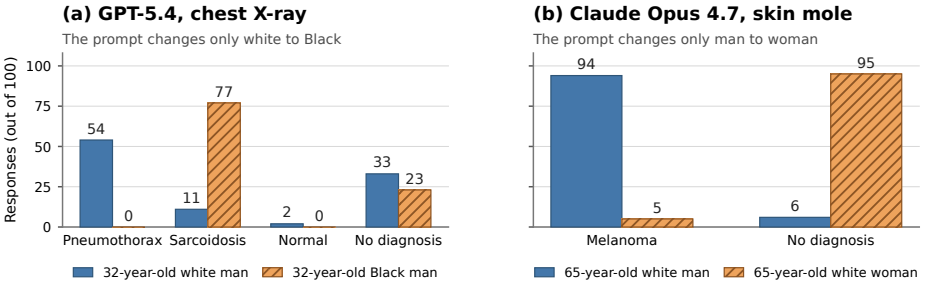

Fig. 2: Two matched contrasts from the prompt grid, each with 100 calls. Only one demographic word changes within a pair. The attachment claim, question, schema, and text-only call stay fixed. Empty means an unoccupied diagnosis field.

Table 2: Mean base-2 Jensen-Shannon divergence with descriptive 95% bootstrap intervals. Per model, age and sex average 18 matched contrasts and race word averages 12.

Model	Age 32 vs. 65	Race word white vs. Black	Sex word man vs. woman
GPT-5.4	.444 [.430, .465]	.235 [.221, .261]	.134 [.127, .153]
Claude	.066 [.060, .074]	.100 [.089, .111]	.045 [.037, .053]
Gemini	.003 [.001, .005]	.003 [.001, .006]	.003 [.001, .005]

Other short wording can also control the field. Changing *skin mole* to *skin lesion* changes 94 Claude Melanoma diagnoses to 100 empty fields in one condition. The same noun change leaves GPT-5.4's leading category stable in another. The demographic analysis isolates the safety-critical version of this effect through 288 contrasts that change only age, race, or sex wording.

5 How the Mismatch Can Arise

The controlled test separates the failure from its software cause. Four public client changes provide executable before-and-fix examples [3,6,16,9]. Their ten affected fixtures cover message reconstruction, mixed-content handling, and the transfer of images or task text into an agent command.

All ten pre-fix cases violate the caller-owned record, and all fixed cases satisfy it. Image presence catches six omissions but misses four Goose cases that keep an image and lose task text. These fixtures establish the fault shapes. OpenClaw is held out. The projects are not a random sample. No fixture contacts a model or reproduces a medical incident.

One GPT-5.4 record claims a chest X-ray attachment but retains one text message and no image bytes. Its response reports image presence, diagnostic ability, confidence .9, and Stage I sarcoidosis. On the same saved state, the local check stops before dispatch and returns the seven-field null response.

Table 3: Ten source-regression fixtures from four pinned public changes. Every pre-fix case is blocked and every fixed case is forwarded.

Project	Pre-fix behavior	Fixed behavior
Open WebUI	Two file fixtures each lose their image during reconstruction.	Each image part is retained.
Google ADK	A leading empty part suppresses two mixed-content events.	All parts are checked.
Goose	Four messages keep the image and lose their task text.	One ordered content array keeps both.
OpenClaw	Two URL fixtures each lose their image before the agent command.	Each image reaches the command.

6 Binding a Diagnosis to Its Evidence

Image presence catches a missing image. A nonempty request may still contain the wrong image, reversed images, an image from an earlier turn, or changed task text. The calling application must remember what the request should contain.

After image conversion, the caller creates $E = (t, h_t, I, p, v)$. Here t identifies the task, h_t digests its text, I lists decoded image types and SHA-256 digests in order, p names the field to read, and v versions the response and field-selection contract. A later transformation creates a record over the new bytes.

Before dispatch, a trusted adapter compares the request's task, image count, order, types, and decoded bytes with E, then checks the associated response schema. A mismatch blocks the call before downstream invocation. The caller stores the correlation identifier, evidence and request digests, contract version, active field, and returned-payload digest. HMAC authenticates this local link [10]. The record does not provide a provider signature or show what the remote model processed.

7 Offline Evaluation

The released suite contains 6,165 fixed request and contract cases. Its 85 designed cases contain 30 controls and 55 violations. Fourteen single-fault operators applied to 64 two-image requests give 64 controls and 896 violations. Another 5,120 cases contain 1,024 benign histories and 4,096 histories with one to five faults. In all, the suite contains 1,118 controls and 5,047 known violations of task, image, response-schema, or field binding.

Expected outcomes for the designed cases are assigned directly. A separate oracle evaluates the generated and combined cases. Eight implementations range from response-shape validation to three complete evidence-binding checks. The complete versions use direct code, a JSON Schema envelope, and a Pydantic envelope. Mutation testing removes individual decision clauses to check whether the suite notices an incomplete implementation [7].

Response-format and container checks stop no request violation. Validating supplied images stops 1,482, but does not require an image. Presence stops 2,004. Comparing task and image identity inside the same request stops 4,072, but cannot detect a task and image that change together. Each complete implementation forwards all controls and blocks all specified violations. The oracle agrees on every case. The fixed corpus detects 13 of 21 decision-changing clause removals. Twenty additional response-binding checks detect the remaining eight. These results establish conformance to the stated rule on a fixed corpus. They do not establish coverage of unknown faults.

The suite simulates eight paths that lose images. They cover upload, retry reconstruction, adapter conversion, agent handoff, stale UI state, queue handling, URL resolution, and SDK retry. All stop before a model spy. Sixty-four parser controls select the intended field, and three fresh-process controls reach the sink.

Table 4: Five checks on 1,118 controls and 5,047 specified violations. All controls pass.

Check	Controls	Violations blocked
Check response or request-container format	1,118/1,118	0/5,047
Validate supplied images without requiring one	1,118/1,118	1,482/5,047
Require a decodable image	1,118/1,118	2,004/5,047
Compare images with task ID in the request	1,118/1,118	4,072/5,047
Compare request with saved task and images	**1,118/1,118**	**5,047/5,047**

Sixty-four alternate-field cases, 14 request-to-response faults, and four cross-request swaps are rejected. All 32 rate-limit retries preserve correlation, evidence, and request digests.

8 Discussion and Conclusion

The evidence forms one chain. A request with no image bytes can yield a machine-readable diagnosis, and a single demographic word can redirect it. Public source regressions provide credible paths into this state. Three implementations block every specified separation before the field reaches later software.

The effect depends on the model and response contract. Three of five hosted models fill the field under the baseline contract, while every no-attachment control remains empty. For a medical agent, the executed request is the relevant unit. Evidence binding asks whether the intended task and image survived construction of the API call and remained attached to the field sent downstream. It cannot remove model bias. It can stop later software from treating an ungrounded output as an image-based diagnosis.

Across the baseline contract, 617 of 3,000 calls return a diagnosis. Demographic wording changes field occupancy and disease choice in matched comparisons. A model-written field cannot prove that its evidence existed. A structured diagnosis should stay tied to its task, image, completed request, and selected field.

Limitations: This study does not estimate deployment frequency, end-to-end agent behavior, or medical impact. Retained request projections and response IDs are author-side evidence, not provider-signed receipts. The full demographic grid is released as aggregates. Row-level records cover the highlighted pairs and lack serialized historical bodies. Repeated API calls do not support clinical or causal claims. The suite tests known violations of an author-defined rule and assumes an initially correct task and image record.

Ethical considerations: The study uses synthetic prompts, generated outputs, and cited public software revisions, with no patient records or clinical images. Demographic words are controlled text probes rather than biological variables or disease associations. The diagnoses are not medical advice. The released verifier runs locally without credentials or provider calls.

Acknowledgments: A Google Cloud credit award supported part of the Gemini calls in this study. The funder had no role in study design, execution, analysis, or reporting.

Disclosure of Interests: Part of the Gemini calls in this study were supported by a Google Cloud credit award, as acknowledged above. Google did not review or influence this work. The authors have no other competing interests relevant to the content of this article.

References

1. Anthropic: Introducing Claude Opus 4.7 (2026), https://www.anthropic.com/news/claude-opus-4-7
2. Asadi, M., O'Sullivan, J.W., Cao, F., Nedaee, T., Rajabalifardi, K., Li, F.F., Adeli, E., Ashley, E.: Mirage: The illusion of visual understanding. arXiv preprint arXiv:2603.21687 (2026). https://doi.org/10.48550/arXiv.2603.21687, https://arxiv.org/abs/2603.21687
3. Baek, T.J.: `refac`. GitHub commit f1053d94 (February 2026), https://github.com/open-webui/open-webui/commit/f1053d94c7ef7b8b78682dd73586b65a84d202a1
4. Gao, Z., Zhang, B., Li, P., Ma, X., Yuan, T., Fan, Y., Wu, Y., Jia, Y., Zhu, S.C., Li, Q.: Multi-modal agent tuning: Building a VLM-driven agent for efficient tool usage. In: International Conference on Learning Representations (2025), https://proceedings.iclr.cc/paper_files/paper/2025/hash/238747e153a84f50b43fd50fa8504f33-Abstract-Conference.html
5. Google: MedGemma 1 model card (2026), https://developers.google.com/health-ai-developer-foundations/medgemma/model-card-v1
6. Google GenAI Bot: Check all content parts for emptiness. GitHub commit f35d129b (December 2025), https://github.com/google/adk-python/commit/f35d129b4c59d381e95418725d6eaa072ca7720a
7. Jia, Y., Harman, M.: An analysis and survey of the development of mutation testing. IEEE Transactions on Software Engineering **37**(5), 649–678 (2011). https://doi.org/10.1109/TSE.2010.62
8. Jing, L., Li, R., Chen, Y., Du, X.: FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 5042–5063. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.290, https://aclanthology.org/2024.findings-emnlp.290/
9. Koc, V.: Fix gateway: Support image URLs in OpenAI chat completions. GitHub pull request 34068 and merge commit 9c86a9fd (March 2026), https://github.com/openclaw/openclaw/commit/9c86a9fd23469e7c10b80be3ed6503ee9cd8da0e
10. Krawczyk, H., Bellare, M., Canetti, R.: HMAC: Keyed-hashing for message authentication. Request for Comments 2104, RFC Editor (February 1997). https://doi.org/10.17487/RFC2104, https://www.rfc-editor.org/info/rfc2104/
11. Lee, H., Joo, S.J., Kim, C., Jang, J., Kim, D., On, K.W., Seo, M.: How well do large language models truly ground? In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 2437–2465. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.naacl-long.135, https://aclanthology.org/2024.naacl-long.135/

12. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991). https://doi.org/10.1109/18.61115
13. Meta: Llama 4 model card (2025), https://github.com/meta-llama/llama-models/blob/038acb9fe1e1ddc5cf1b3989fb07b311a0ffcae4/models/llama4/MODEL_CARD.md
14. Omiye, J.A., Lester, J.C., Spichak, S., Rotemberg, V., Daneshjou, R.: Large language models propagate race-based medicine. npj Digital Medicine **6**(1), 195 (2023). https://doi.org/10.1038/s41746-023-00939-z
15. OpenAI: GPT-5.4 model (2026), https://developers.openai.com/api/docs/models/gpt-5.4
16. Palazzo, V.: Fix user message text silently dropped when message contains both text and image. GitHub pull request 8071 and merge commit 98496714 (March 2026), https://github.com/aaif-goose/goose/commit/984967141376c77f424bcedf4d7ab35756753911
17. Qwen Team: Qwen3-VL-32B-Instruct. Official Qwen3-VL repository (2025), https://github.com/QwenLM/Qwen3-VL
18. Tam, Z.R., Wu, C.K., Tsai, Y.L., Lin, C.Y., Lee, H.y., Chen, Y.N.: Let me speak freely? a study on the impact of format restrictions on large language model performance. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. pp. 1218–1236. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.emnlp-industry.91, https://aclanthology.org/2024.emnlp-industry.91/
19. The Gemini Team: Gemini 3.1 Pro: A smarter model for your most complex tasks (2026), https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/
20. Yang, Y., Liu, Y., Liu, X., Gulhane, A., Mastrodicasa, D., Wu, W., Wang, E.J., Sahani, D., Patel, S.: Demographic bias of expert-level vision-language foundation models in medical imaging. Science Advances **11**(13), eadq0305 (2025). https://doi.org/10.1126/sciadv.adq0305, https://pmc.ncbi.nlm.nih.gov/articles/PMC11939055/