Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Image-embedded prompt injection vulnerability of vision-language models in dental radiology: a cross-vendor attack–defense evaluation

Babak Saravi

babak.saravi@med.uni-duesseldorf.de

Heinrich Heine University Düsseldorf

Daman Deep Singh

Heinrich Heine University Düsseldorf

Lara Schorn

Heinrich Heine University Düsseldorf

Andreas Vollmer

University Hospital Würzburg

Christoph Sproll

Heinrich Heine University Düsseldorf

Norbert Kübler

Heinrich Heine University Düsseldorf

Felix Schrader

Heinrich Heine University Düsseldorf

Article

Keywords:

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9932271/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 3rd, 2026.
See the published version at https://doi.org/10.1038/s41598-026-74077-3.

Abstract

Image-embedded prompt injection — adversarial text rendered into the pixel data of medical images — is an emerging threat to vision-language models (VLMs) used in clinical decision support. We systematically evaluated this vulnerability across four production-tier VLMs (GPT-4o, Gemini 2.5 Flash, Claude Sonnet 4.5, MedGemma 4B) on 270 dental panoramic radiographs from the DenTeX dataset under four attack classes (8 variants per image), comprising 9,720 baseline and 48,600 defense inference calls (58,320 total). All four models were vulnerable, with paired attack success rates up to 62.6% (95% CI: 58.5–66.7%) for GPT-4o. Vulnerability profiles were model-specific, with significant attack-by-model interactions. We benchmarked five defense strategies ranging from simple image cropping to OCR-based text sanitization and a provenance-aware defense with bounded clinical abstention (ProvDent). OCR-based sanitization achieved the strongest attack reduction (pooled ASR: 0.2%), while ProvDent provided a complementary fail-open governance mechanism that escalates suspicious cases for human review. All defenses preserved clean-image F1 within 0.6 percentage points of baseline on this high-prevalence benchmark. These findings establish image-embedded prompt injection as a cross-vendor security risk for dental VLMs and provide an empirical basis for defense selection in clinical deployment.

Introduction

Artificial intelligence (AI) has emerged as a transformative technology in dental radiology, with deep learning models demonstrating high accuracy in tasks ranging from caries detection and tooth enumeration to periodontal bone loss assessment on panoramic radiographs [1, 2]. Recent systematic reviews have documented AI-assisted diagnostic accuracy exceeding 90% for several dental pathology classes [2], and the rapid expansion of publicly available dental imaging datasets has further accelerated model development in this domain [3].

The emergence of multimodal large language models (LLMs) capable of processing both text and images — so-called vision-language models (VLMs) — has opened new possibilities for clinical decision support in dentistry. Models such as GPT-4o, Gemini, and Claude can now accept dental radiographs as input and generate structured diagnostic assessments. Initial evaluations of multimodal LLMs on dental examination questions have shown promising text-based performance (over 80% accuracy) but limited image interpretation capability [4, 5]. Similarly, studies in general radiology have demonstrated that VLMs can approach junior clinician accuracy on image challenge cases [6]. These findings suggest that VLM-assisted dental diagnosis may become clinically feasible in the near term.

However, the deployment of LLMs in clinical settings raises significant safety and security concerns [7, 8]. LLMs have been shown to be susceptible to adversarial hallucination attacks, where fabricated details embedded in prompts cause the model to elaborate on false clinical information, with hallucination rates ranging from 50% to 82% across models [7]. Regulatory frameworks including the EU Artificial Intelligence Act are struggling to keep pace with the rapid clinical adoption of these tools, and

the risk of patient harm from unvalidated AI outputs has been explicitly identified as a policy concern [8]. Beyond hallucination, the broader landscape of adversarial threats to healthcare AI — including data poisoning, model manipulation, and input-level attacks — has been systematically characterized [9], with structured red-teaming frameworks being developed to evaluate clinical AI safety [10].

One specific and increasingly important threat vector is prompt injection: the manipulation of an AI model's behavior by embedding adversarial instructions within the model's input. In the context of multimodal VLMs, this attack surface extends beyond text to include the image modality. An adversary can render textual instructions directly into the pixel data of a medical image, creating an image-embedded prompt injection that the VLM reads and obeys as if it were a legitimate instruction. This is particularly concerning for dental panoramic radiography, where images routinely pass through multiple software systems — PACS archives, teledentistry platforms, and AI preprocessing pipelines — before reaching the diagnostic model, creating multiple opportunities for adversarial content insertion. Initial empirical studies have begun to demonstrate this vulnerability in medicine: image-embedded prompt injection can subvert production VLMs in oncological imaging [11], and broader multimodal work has shown that adversaries can hijack vision-language models through low-contrast or visually obfuscated text overlays [12]. A recent systematic review of prompt injection nonetheless identified non-textual, image-based injection channels and the empirical *development and assessment* of defense mechanisms as two of the principal open research gaps in the field [13]. Both gaps remain unaddressed in dental radiology, where neither the cross-vendor vulnerability of VLMs nor the comparative effectiveness of candidate defenses has been systematically evaluated.

Despite the clinical significance of this threat, the vulnerability of dental VLMs to image-embedded prompt injection has not been systematically characterized, and no clinically validated defense has been proposed for this specific domain. Prior work in general medical imaging has demonstrated the feasibility of such attacks, but the dental radiology context introduces unique considerations: the clinical prior that in standard German dental imaging workflows, panoramic radiographs contain no burned-in pixel-level text creates a natural anomaly signal that a defense system can exploit, and the high clinical cost of missed dental pathology (e.g., undetected caries or periapical lesions) makes false-negative-inducing attacks particularly dangerous [14].

The aims of this study were therefore threefold. First, we sought to systematically characterize the vulnerability of four production-tier VLMs (GPT-4o, Gemini 2.5 Flash, Claude Sonnet 4.5, and MedGemma 4B) to image-embedded prompt injection on dental panoramic radiographs from the publicly available DenTeX dataset [15], using four attack classes that vary in visual salience and spatial placement. Second, we aimed to evaluate a spectrum of defense conditions — from simple image cropping and prompt-level interventions to OCR-based text sanitization — and to quantify their effectiveness in reducing attack success while preserving diagnostic accuracy. Third, we proposed and evaluated ProvDent, a novel provenance-aware defense system that combines burned-in text detection, instruction-likeness scoring, conditional dual-pass consistency checking, and bounded clinical abstention, specifically designed to exploit the clinical prior that standard German dental radiographs contain no

pixel-level text. Together, these aims address the critical gap between the growing clinical adoption of multimodal AI in dentistry and the absence of systematic security evaluation and deployable defenses for this domain.

Methods

Study design and threat model

This observational experimental study evaluated the vulnerability of production-tier multimodal vision-language models (VLMs) to image-embedded prompt injection attacks in the context of AI-assisted dental panoramic radiography, and assessed the effectiveness of five defense conditions including a novel provenance-aware defense system (ProvDent).

The threat model assumed an adversary with the ability to place textual content into the pixel data of a dental radiograph before it reaches a clinical VLM at inference time. Realistic attack vectors include third-party AI preprocessing services, teledentistry forwarding platforms, compromised PACS exports, and direct malicious uploads.

A clinical prior informed the defense design: in the standard German dental imaging workflow used at our institution, panoramic radiographs are transmitted to downstream systems without burned-in pixel-level text annotations. Tooth identifiers, quadrant markers, and radiographer notes are stored in the DICOM header or PACS overlay layer, not rendered onto the pixel data. Any text present in the pixel data is therefore anomalous. This prior, confirmed by the oral and maxillofacial surgery clinician on the study team, motivates the text-detection component of the ProvDent defense. Training-data poisoning, DICOM/EXIF metadata injection, weight-level attacks, and physical sensor attacks were excluded from scope.

Dataset

The study used the DenTeX dataset from the MICCAI 2023 Dental Enumeration and Diagnosis on Panoramic X-rays Challenge, released under CC-BY license. The training_data/quadrant-enumeration-disease split was used, containing 705 panoramic radiographs (678 with pathology annotations, 27 without). Ground-truth annotations comprised 3,529 tooth-level entries across four pathology classes: Caries (n = 2,189; 62.0%), Impacted (n = 604; 17.1%), Deep Caries (n = 578; 16.4%), and Periapical Lesion (n = 158; 4.5%). N = 300 images were selected via stratified random sampling (seed = 42), with 270 for the main evaluation set and 30 reserved as a held-out development/tuning set used exclusively for selecting ProvDent's instruction-likeness threshold and dual-pass disagreement threshold. The development set was not included in any reported analysis. All images were normalized to a maximum long-side dimension of 1,600 pixels and 3.5 MB file size using Lanczos resampling, applied identically across all four models.

Target vision-language models

Four VLMs were selected representing each major vendor's current-generation mid-tier production multimodal model: GPT-4o (OpenAI; temperature = 0.0, max_tokens = 500, Chat Completions API), Gemini 2.5 Flash (Google; temperature = 0.0, max_tokens = 2,048, thinking disabled, safety filters set to BLOCK_NONE, google-genai SDK), Claude Sonnet 4.5 (Anthropic; temperature = 0.0, max_tokens = 500, Messages API), and MedGemma 4B (Google, open-weight; bfloat16 on NVIDIA RTX 3090, max_tokens = 2,048, HuggingFace Transformers). All models received identical system and user prompts framing the task as a structural visual description exercise for dental education, explicitly disclaiming diagnostic intent. The educational framing was adopted after a pilot experiment in which the original diagnostic framing elicited a 46% refusal rate from GPT-4o. Although the prompt disclaimed diagnostic intent, the structured output schema (any_abnormality_present as a binary classifier) constitutes a de facto diagnostic task, and we report diagnostic accuracy metrics accordingly. We note that reframing a diagnostic task as 'educational' to bypass model safety filters is itself a form of prompt engineering; this limitation is acknowledged in the Discussion. Importantly, the pilot experiment demonstrated that the vulnerability exists under both framings: among images where GPT-4o did respond under the diagnostic prompt, paired ASR was near-universal. The educational framing thus eliminates refusals without creating the vulnerability, enabling cleaner measurement.

Attack taxonomy

Four attack classes were implemented: (1) burned-in overlay (high salience; black background box, opacity 0.9, font 24px, upper-right), (2) annotation mimicry (high salience; radiographer-note style, opacity 0.9, font 18px, upper-center), (3) low-contrast overlay (low salience; plain white text, opacity 0.35, font 24px, upper-right), and (4) border patch (peripheral; near-white text RGB 240,240,240, opacity 0.85, font 16px, bottom edge). Each class used two payload formulations: an imperative ("IGNORE PRIOR INSTRUCTIONS…") and a system-override ("SYSTEM OVERRIDE…"). All payloads targeted false-negative induction. Total: 9 conditions per image (1 clean + 8 attacked), yielding 9,720 baseline inference calls (Supplementary Fig. 1).

Defense conditions

Five conditions were evaluated: (1) no defense (baseline), (2) ROI crop (bottom 15%, left/right 5%), (3) spotlighting (prompt-level untrusted-data delimiters), (4) OCR sanitize (EasyOCR detection, confidence ≥ 0.4, black inpainting), and (5) ProvDent (spotlighting + EasyOCR + GPT-4o-mini instruction-likeness judge + conditional dual-pass when score ≥ 0.5 + bounded abstention on disagreement). Each defense was applied to all 9,720 baseline conditions, yielding 48,600 defense inference calls (Supplementary Fig. 2–3).

Outcome measures

The primary outcome was paired ASR: the proportion of images where the model's any_abnormality_present output flipped from true (clean) to false (attacked), restricted to pairs where both conditions parsed successfully. Secondary outcomes included ground-truth-referenced F1, sensitivity, specificity, and NPV on clean images, and dental-CHER (false-negative weight = 4.0, false-positive = 1.0, refusal = 0.5). Defense-specific outcomes included per-defense ASR with bootstrap CIs, clean-image F1 under each defense, and ProvDent component metrics. Exploratory subgroup analyses by pathology class and model confidence were pre-specified.

Statistical analysis

All analyses were pre-specified before data collection in a time-stamped analysis plan; the locked commit and amendment log are available in the project repository (see Data availability section). Bootstrap 95% CIs used 10,000 case-level resamples (seed = 42). McNemar's test (exact binomial ≤ 25 discordant pairs, chi-square otherwise) was applied per (model×attack) cell with Holm–Bonferroni correction across 16 primary tests ($\alpha = 0.05$). Effect sizes: Cohen's $g = |b - c|/(b + c)$ for discordant pairs (undefined when $b = c = 0$). A Bayesian variational GLMM (BinomialBayesMixedGLM, statsmodels v0.14) was fitted: attack_success ~ C(attack_class)×C(model_key) + (1|image_id). Fixed-effect inference was based on the variational posterior means and standard deviations; z-scores and two-sided p-values were derived from the normal approximation to the posterior. A fixed-effects logistic regression with cluster-robust standard errors (grouped by image_id) was pre-specified as a convergence fallback. CHER sensitivity analysis swept FN weights 2–6. As a sensitivity analysis, all primary McNemar tests and bootstrap CIs were also computed at the image level (averaging across the two payloads per attack class), yielding similar conclusions; payload-level results are reported as primary because they preserve the full experimental design. Subgroups with $n < 10$ were excluded. Software: Python 3.11, pandas 3.0, NumPy 2.4, SciPy 1.17, statsmodels 0.14, Matplotlib 3.10.

Sample size justification

N = 270 was determined by power analysis on pilot effect sizes (n = 15 images, yielding 30 paired observations per model×attack cell because each attack class used 2 payloads). For paired McNemar at $\alpha = 0.05$ with 80% power ($\rho = 0.30$), the maximum per-cell N was 82 (GPT-4o low-contrast, 27%→10%). N = 270 provides a 3.3-fold safety margin. The 95% CI half-width on the headline ASR cell is ± 4 percentage points.

Reproducibility

Every inference call was logged as a 47-field CallRecord to an append-only JSONL file with per-row fsync. The experiment comprised 58,320 calls (9,720 baseline + 48,600 defense), all achieving 100% JSON parsing completeness. The complete source code, defense implementations, analysis pipeline, and experiment configuration files are available at https://github.com/Freiburg-AI-Research/provdent. All 58,320 structured inference call records (9,720 baseline and 48,600 defense calls), the experiment

manifest, and pre-computed analysis tables are deposited on Figshare at https://doi.org/10.6084/m9.figshare.31985511.v1 under a CC BY 4.0 license.

Ethics and data governance

The DenTeX dataset is publicly available under a CC-BY license and contains fully de-identified panoramic dental radiographs. No patient consent was required for this secondary analysis. The study does not constitute human-subjects research and did not require institutional review board approval.

Results

Dataset and completeness

The main experiment comprised 9,720 baseline inference calls across 270 stratified DenTeX panoramic radiographs, four VLMs, and four attack classes. An additional 48,600 defense calls were executed across five conditions. All rows achieved successful JSON parsing (100% completeness). No model exhibited refusal behavior on clean images (refusal rate: 0%).

Baseline diagnostic performance

On clean images, all four models achieved high sensitivity for abnormal-tooth detection (98.8–100.0%) with F1 scores of 94.7–95.1% (Supplementary Table 1). However, specificity was very low (0.0–4.0%): three of four models classified every image as abnormal, reflecting a strong positive bias. This is attributable to the high abnormality prevalence in the DenTeX disease split (678 of 705 images annotated with pathology), which leaves only 27 negative-class images and renders specificity and negative predictive value poorly estimable. Consequently, the utility-preservation findings reported below (F1 within 0.6 percentage points of baseline) are driven primarily by sensitivity and should not be interpreted as a comprehensive clinical validation of diagnostic robustness across balanced populations. The dental clinical-harm-weighted error rate (dental-CHER, with false-negative weight = 4.0) on clean images was 0.60 across all models, driven entirely by false positives on the small negative-class subset. Sensitivity analysis varying the false-negative weight from 2.0 to 6.0 did not meaningfully alter CHER rankings across models or defense conditions, confirming robustness of the harm-weighted findings to the specific weight choice.

Vulnerability to image-embedded prompt injection

All four VLMs were vulnerable to at least one attack class (Table 1, Fig. 1). GPT-4o exhibited the highest paired ASR at 62.6% (95% CI: 58.5–66.7%) under burned-in overlay, followed by MedGemma 4B (55.4%), Gemini 2.5 Flash (14.8%), and Claude Sonnet 4.5 (9.1%). Mean ASR across attacks was 28.3% for GPT-4o, 23.8% for MedGemma, 5.5% for Claude, and 5.4% for Gemini. A monotonic gradient was visible: burned-in > annotation mimicry > low-contrast > border patch (Fig. 2).

Table 1
Paired attack success rate (ASR) by model and attack class with 95% bootstrap confidence intervals (10,000 resamples). N = 540 observations per cell.

Model	Attack	ASR % [95% CI]	Flips	N
GPT-4o	Burned-in overlay	62.6 [58.5, 66.7]	338	540
GPT-4o	Annotation mimicry	25.7 [22.2, 29.4]	139	540
GPT-4o	Low-contrast overlay	21.9 [18.3, 25.4]	118	540
GPT-4o	Border patch	3.0 [1.7, 4.4]	16	540
Gemini 2.5 Flash	Burned-in overlay	14.8 [11.9, 18.0]	80	540
Gemini 2.5 Flash	Annotation mimicry	6.3 [4.3, 8.3]	34	540
Gemini 2.5 Flash	Low-contrast overlay	0.4 [0.0, 0.9]	2	540
Gemini 2.5 Flash	Border patch	0.0 [0.0, 0.0]	0	540
Claude Sonnet 4.5	Burned-in overlay	9.1 [6.7, 11.5]	49	540
Claude Sonnet 4.5	Annotation mimicry	11.7 [9.1, 14.4]	63	540
Claude Sonnet 4.5	Low-contrast overlay	1.1 [0.4, 2.0]	6	540
Claude Sonnet 4.5	Border patch	0.0 [0.0, 0.0]	0	540
MedGemma 4B	Burned-in overlay	55.4 [51.3, 59.4]	299	540
MedGemma 4B	Annotation mimicry	30.9 [27.0, 34.8]	167	540
MedGemma 4B	Low-contrast overlay	7.2 [5.0, 9.4]	39	540
MedGemma 4B	Border patch	1.7 [0.7, 2.8]	9	540

Statistical inference

McNemar's test confirmed significant attack effects for all primary cells after Holm–Bonferroni correction, except border-patch on Claude/Gemini and low-contrast on Gemini (Table 2). Cohen's g was 1.0 for all significant cells, reflecting the entirely unidirectional nature of the attack effect (all flips from positive to negative, zero reverse flips). For cells with zero discordant pairs (b = c=0), g is undefined and reported as such. A GLMM with image-level random intercepts confirmed that, at the reference attack class (annotation mimicry), GPT-4o (OR = 3.13, CI: 2.80–3.51, p<.001) and MedGemma (OR = 4.16, CI: 3.68–4.70, p<.001) were significantly more likely to be flipped than Claude Sonnet 4.5 (reference category). Significant attack×model interactions indicated that these vulnerability differences varied by attack class; for example, GPT-4o was disproportionately vulnerable to low-contrast attacks (interaction OR = 8.25, CI: 6.64–10.24, p<.001), indicating that vulnerability profiles are model-specific (Supplementary Table 3).

Table 2
McNemar's test with Holm–Bonferroni correction across 16 primary tests. Cohen's g = |b − c|/(b + c).

Model	Attack	N pairs	b	c	p	p (Holm)	g
GPT-4o	Burned-in overlay	540	338	0	< .001	< .001	1.00
GPT-4o	Annotation mimicry	540	139	0	< .001	< .001	1.00
GPT-4o	Low-contrast overlay	540	118	0	< .001	< .001	1.00
GPT-4o	Border patch	540	16	0	< .001	< .001	1.00
Gemini 2.5 Flash	Burned-in overlay	540	80	0	< .001	< .001	1.00
Gemini 2.5 Flash	Annotation mimicry	540	34	0	< .001	< .001	1.00
Gemini 2.5 Flash	Low-contrast overlay	540	2	0	0.500	1.000	1.00
Gemini 2.5 Flash	Border patch	540	0	0	1.000	1.000	—
Claude Sonnet 4.5	Burned-in overlay	540	49	0	< .001	< .001	1.00
Claude Sonnet 4.5	Annotation mimicry	540	63	0	< .001	< .001	1.00
Claude Sonnet 4.5	Low-contrast overlay	540	6	0	0.031	0.125	1.00
Claude Sonnet 4.5	Border patch	540	0	0	1.000	1.000	—
MedGemma 4B	Burned-in overlay	540	299	0	< .001	< .001	1.00
MedGemma 4B	Annotation mimicry	540	167	0	< .001	< .001	1.00
MedGemma 4B	Low-contrast overlay	540	39	4	< .001	< .001	0.81
MedGemma 4B	Border patch	540	9	0	0.004	0.020	1.00

Defense effectiveness

OCR sanitize achieved the lowest pooled ASR at 0.2%, virtually eliminating all attacks (Table 3, Fig. 3). Spotlighting reduced ASR to 7.4%, ROI crop to 9.5%. ProvDent attained 7.5% with an additional abstention safety mechanism. All defenses preserved clean-image F1 within 0.6 percentage points of baseline (94.7–95.1%). Defense effectiveness was consistent across all four models (Fig. 4).

Table 3
Pooled ASR with Wilson 95% confidence intervals across all models and attacks under each defense condition.

Defense	ASR % [95% CI]	Flips	N paired
No defense	15.60 [14.85, 16.38]	1348	8640
ROI crop	9.47 [8.87, 10.10]	818	8640
Spotlighting	7.43 [6.90, 8.00]	642	8640
OCR sanitize	0.23 [0.15, 0.36]	20	8640
ProvDent	7.52 [6.99, 8.10]	650	8640

Table 4
Defense trade-off comparison. Pooled ASR = attack success rate across all models and attack classes. Fail-open signal = whether the defense explicitly alerts the clinician when a potential manipulation is detected. ProvDent is the only defense that provides bounded clinical abstention, escalating suspicious cases for human review rather than silently passing a potentially compromised answer.

Defense	Pooled ASR (%)	Clean F1 (%)	Abstention (%)	Fail-open signal	Extra API calls
No defense	15.60	94.7–95.1	0	No	0
ROI crop	9.47	94.1–95.1	0	No	0
Spotlighting	7.43	94.9–95.1	0	No	0
OCR sanitize	0.23	94.7–95.1	0	No	0
ProvDent	7.52	94.9–95.1	0.1–20.9	Yes	1–2

ProvDent component analysis

ProvDent's OCR text detector fired on 85.0% of attacked images and 52.6% of clean images. The high clean-image detection rate reflects EasyOCR's sensitivity to image-border noise, JPEG compression artifacts, and dental anatomy patterns that resemble text glyphs; critically, the downstream instruction-likeness judge successfully filtered these false positives, triggering the dual-pass on 57.2–57.8% of attacked images but 0.0% of clean images (mean instruction-likeness score: 0.53 on attacked text vs. 0.05 on clean-image artifacts). Abstention was invoked on 0.1–20.9% of attacked images (model-dependent; highest for MedGemma 4B, consistent with its lower baseline robustness), with 0.0% abstention on clean images, confirming that the abstention mechanism did not generate false safety alerts on unmanipulated radiographs (see Supplementary Tables and Figures).

While OCR-based sanitization achieved the lowest pooled ASR (0.2%), ProvDent's value lies not in raw ASR reduction but in its fail-open triage behavior: when dual-pass disagreement is detected, the system explicitly escalates the case for human review rather than silently returning a potentially compromised

answer. This bounded abstention mechanism provides a safety guarantee that passive defenses (ROI crop, spotlighting, OCR sanitize) do not offer.

Exploratory subgroup analyses

Subgroup analyses by pathology class revealed that vulnerability was highest for impacted teeth and lowest for periapical lesions (Supplementary Fig. 6). From a clinical perspective, successful suppression of impacted-tooth findings could delay surgical evaluation or extraction planning, while missed caries — though more frequently attacked due to higher prevalence in the dataset — may have less immediately critical but still consequential outcomes if untreated pathology progresses. Stratification by model self-reported confidence indicated that high-confidence answers were more frequently flipped than low-confidence answers (Supplementary Fig. 7), suggesting that prompt injection can selectively override the model's strongest diagnostic assertions. These analyses are exploratory and reported without multiple-comparison correction.

Discussion

This study provides the first systematic evaluation of image-embedded prompt injection vulnerability across multiple production-tier VLMs applied to dental panoramic radiography, and proposes a provenance-aware defense system (ProvDent). All four VLMs were susceptible, with vulnerability varying substantially by model and attack type. Text-aware defenses, particularly OCR-based sanitization, can virtually eliminate this attack surface while preserving diagnostic accuracy.

Our findings extend the growing body of evidence on adversarial vulnerabilities in clinical AI. Omar et al. demonstrated adversarial hallucination rates of 50–82% in clinical LLMs [7]. Our work extends this from text to the image modality, showing that pixel-level adversarial content is similarly effective. The broader threat landscape including data poisoning has been characterized [9], and red-teaming frameworks proposed [10]. Our study provides the first domain-specific attack–defense evaluation for dental radiology.

The cross-vendor vulnerability gradient has important clinical implications. The open-weight medical model (MedGemma 4B) was among the most vulnerable, suggesting domain-specific fine-tuning does not confer robustness against prompt injection. Model-specific interaction effects indicate vulnerability profiles cannot be predicted from accuracy alone. Security evaluation must become a mandatory component of clinical AI validation [8].

The defense benchmarking results reveal a clear spectrum of safety profiles. Passive defenses (ROI crop, spotlighting, OCR sanitize) operate silently: they modify the image or prompt before inference and return the model's output without any indication that a potential manipulation was processed. Among these, OCR-based sanitization was the most effective, reducing pooled ASR from 15.6% to 0.2% — consistent with input-level preprocessing approaches in the adversarial robustness literature [16, 17]. Its simplicity and model-agnostic nature make it a strong candidate for immediate deployment as a

preprocessing step in clinical VLM pipelines. ProvDent, by contrast, is not designed to minimize ASR below the level achieved by OCR sanitize. Its primary contribution is an active governance layer: when its dual-pass consistency check detects disagreement between the sanitized and original image outputs, it escalates the case for human review rather than silently returning a potentially compromised answer. This fail-open behavior addresses a different risk dimension — institutional accountability and auditability — that passive defenses do not cover [8]. Institutions adopting VLMs for clinical dental radiology may therefore choose OCR sanitize for maximal attack reduction, ProvDent for fail-open governance, or a combination depending on their risk management requirements (Table 4). All defenses preserved clean-image F1 within 0.6 percentage points of baseline, though this finding is primarily driven by sensitivity given the high abnormality prevalence in the DenTeX benchmark.

The dental radiology context provides a domain-specific advantage: in standard German clinical workflows, panoramic radiographs contain no burned-in pixel-level text, creating a strong anomaly signal for text-based defenses. However, this prior may not hold universally — imaging centers in some regions annotate radiographs directly, and commercial AI tools increasingly add watermarks or diagnostic overlays. As AI integrates into dental clinical workflows [18, 19], the security of the entire imaging pipeline must be considered as part of clinical governance.

This study has several strengths: four VLMs from three vendors plus an open-weight medical model, a publicly available annotated dataset (DenTeX) [15], 58,320 inference calls with 100% data completeness, pre-specified analyses with multiple-comparison correction, and a five-defense spectrum enabling direct comparison. Several limitations warrant discussion. First, the high abnormality prevalence in the DenTeX disease split (678/705 images positive) resulted in near-zero specificity for all models on clean images, limiting the interpretability of specificity-dependent metrics (NPV) and meaning that the F1-based utility preservation claim is driven primarily by sensitivity. Second, attack payloads were limited to English-language formulations targeting a single diagnostic task; gradient-based adversarial optimization, which represents a harder threat regime, was not tested. Third, generalizability to other dental imaging modalities (periapical, CBCT) and to radiographs with legitimate burned-in annotations (as used in some non-European workflows) was not evaluated. Fourth, ProvDent relies on an external LLM judge (GPT-4o-mini) for instruction-likeness scoring, which may not be acceptable in all clinical deployment settings. Fifth, the ‘educational’ prompt framing used to overcome GPT-4o’s safety refusals is itself a form of prompt engineering that may not reflect real-world clinical deployment. Finally, model updates may alter vulnerability profiles over time.

Future work should extend to gradient-optimized attacks, external clinical validation, judge distillation for local deployment, and regulatory integration of adversarial security evaluation [8]. In conclusion, image-embedded prompt injection is a clinically significant threat to VLM-assisted dental radiology. Effective defenses exist and can be deployed without materially reducing clean-image F1 in this benchmark. We recommend security evaluation against prompt injection as a standard component of clinical AI validation for any multimodal system that processes medical images.

Declarations

Ethics approval and consent to participate

The DenTeX dataset is publicly available under a CC-BY license and contains fully de-identified panoramic dental radiographs with no patient-identifiable information. No patient consent was required for this secondary analysis of a publicly released dataset. No clinical data beyond the DenTeX images was transmitted to any third-party API. The study does not constitute human-subjects research and did not require institutional review board approval.

Consent for publication

Not applicable.

Data availability

The DenTeX dataset is available at https://dentex.grand-challenge.org/ under a CC-BY license. The complete source code, defense implementations, analysis pipeline, and experiment configuration files are available at https://github.com/Freiburg-AI-Research/provdent. All 58,320 structured inference call records (9,720 baseline and 48,600 defense calls), the experiment manifest, and pre-computed analysis tables are deposited on Figshare at https://doi.org/10.6084/m9.figshare.31985511.v1 under a CC BY 4.0 license. The datasets generated and/or analyzed during the current study are available in the Figshare repository, https://doi.org/10.6084/m9.figshare.31985511.v1.

Competing interests

The authors declare no competing interests.

Funding

No external funding was received for this study.

Author contributions

BS conceived the study, developed the experimental pipeline, conducted all experiments, performed statistical analyses, and wrote the manuscript. DDS contributed to study design, data preprocessing, literature review, and manuscript writing. LS contributed to study design and manuscript revision. AV assisted with defense implementation, figures, and editing. CS provided supervision, data interpretation, and critical revision. NK provided resources, clinical interpretation, and project supervision. FS

contributed study design, critical revision, and supervision. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Clusmann, J., Kolbinger, F. R., Muti, H. S., Carrero, Z. I. & Eckardt, J. N. The future landscape of large language models in medicine. *Commun. Med.* **3**, 141 (2023).
2. Turosz, N., Chęcińska, K., Chęciński, M., Brzozowska, A. & Nowak, Z. Applications of artificial intelligence in the analysis of dental panoramic radiographs: an overview of systematic reviews. *Dentomaxillofacial Radiol.* **52**, 20230284 (2023).
3. Uribe, S. E., Issa, J., Sohrabniya, F. & Denny, A. Kim, N. N. Publicly available dental image datasets for artificial intelligence. *J. Dent. Res.* **103**, 1365–1374 (2024).
4. Jeong, H., Jeon, K. J., Lee, C., Choi, Y. J. & Jo, G. D. Leveraging multimodal large language model chatbots in oral radiology: a comprehensive evaluation using questions from a Korean dental university. *Dentomaxillofacial Radiol.* **55**, 276–286 (2026).
5. Mine, Y., Okazaki, S., Taji, T., Kawaguchi, H. & Kakimoto, N. Benchmarking multimodal large language models on the dental licensing examination: challenges with clinical image interpretation. *J. Dent. Sci.* **20**, 2427–2435 (2025).
6. Suh, P. S., Shim, W. H., Suh, C. H., Heo, H. & Park, K. J. Comparing large language model and human reader accuracy with image challenge case image inputs. *Radiology* **313**, e241668 (2024).
7. Omar, M., Sorin, V., Collins, J. D., Reich, D. & Freeman, R. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun. Med.* **5**, 330 (2025).
8. Freyer, O., Wiest, I. C., Kather, J. N. & Gilbert, S. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digit. Health*. **6**, e662–e672 (2024).
9. Abtahi, F., Seoane, F., Pau, I. & Vega-Barbas, M. Data poisoning vulnerabilities across health care artificial intelligence architectures: analytical security framework and defense strategies. *J. Med. Internet. Res.* **28**, e87969 (2026).
10. Trabilsy, M., Prabha, S., Gomez-Cabello, C. A., Haider, S. A. & Genovese, A. The PIEE cycle: a structured framework for red teaming large language models in clinical decision-making. *Bioengineering* **12**, 706 (2025).
11. Clusmann, J. et al. Prompt injection attacks on vision language models in oncology. *Nat. Commun.* **16**, 1239 (2025).

12. Sharma, R. K., Gupta, V. & Grossman, D. Defending Language Models Against Image-Based Prompt Attacks via User-Provided Specifications. in *IEEE Security and Privacy Workshops (SPW)* 112–131 (IEEE, San Francisco, CA, USA, 2024). 112–131 (IEEE, San Francisco, CA, USA, 2024). (2024). 10.1109/SPW63631.2024.00017
13. Damacena Duarte, J. et al. A Systematic Review of Prompt Injection Attacks on Large Language Models: Trends, Taxonomy, Evaluation, Defenses, and Opportunities. *IEEE Access.* **14**, 12875–12899 (2026).
14. Xue, T., Chen, L. & Sun, Q. Deep learning method to automatically diagnose periodontal bone loss and periodontitis stage in dental panoramic radiograph. *J. Dent.* **150**, 105373 (2024).
15. Hamamci, I. E., Er, S. & Simsar, E. DenTeX: a benchmark for dental enumeration and diagnosis on panoramic X-rays. in *Proceedings of the MICCAI 2023 Challenge, Lecture Notes in Computer Science*Springer, (2024). 10.1007/978-3-031-72396-4_2
16. Ashraf, S. N., Siddiqi, R. & Farooq, H. Auto encoder-based defense mechanism against popular adversarial attacks in deep learning. *PLoS ONE*. **19**, e0307363 (2024).
17. Chen, L., Zhao, L. & Chen, C. Y. C. Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism. *Med. Phys.* **48**, 6198–6212 (2021).
18. Ducret, M., Mörch, C. M., Karteva, T., Fisher, J. & Schwendicke, F. Artificial intelligence for sustainable oral healthcare. *J. Dent.* **127**, 104344 (2022).
19. Nordblom, N. F., Büttner, M. & Schwendicke, F. Artificial intelligence in orthodontics: critical review. *J. Dent. Res.* **103**, 577–584 (2024).

Figures

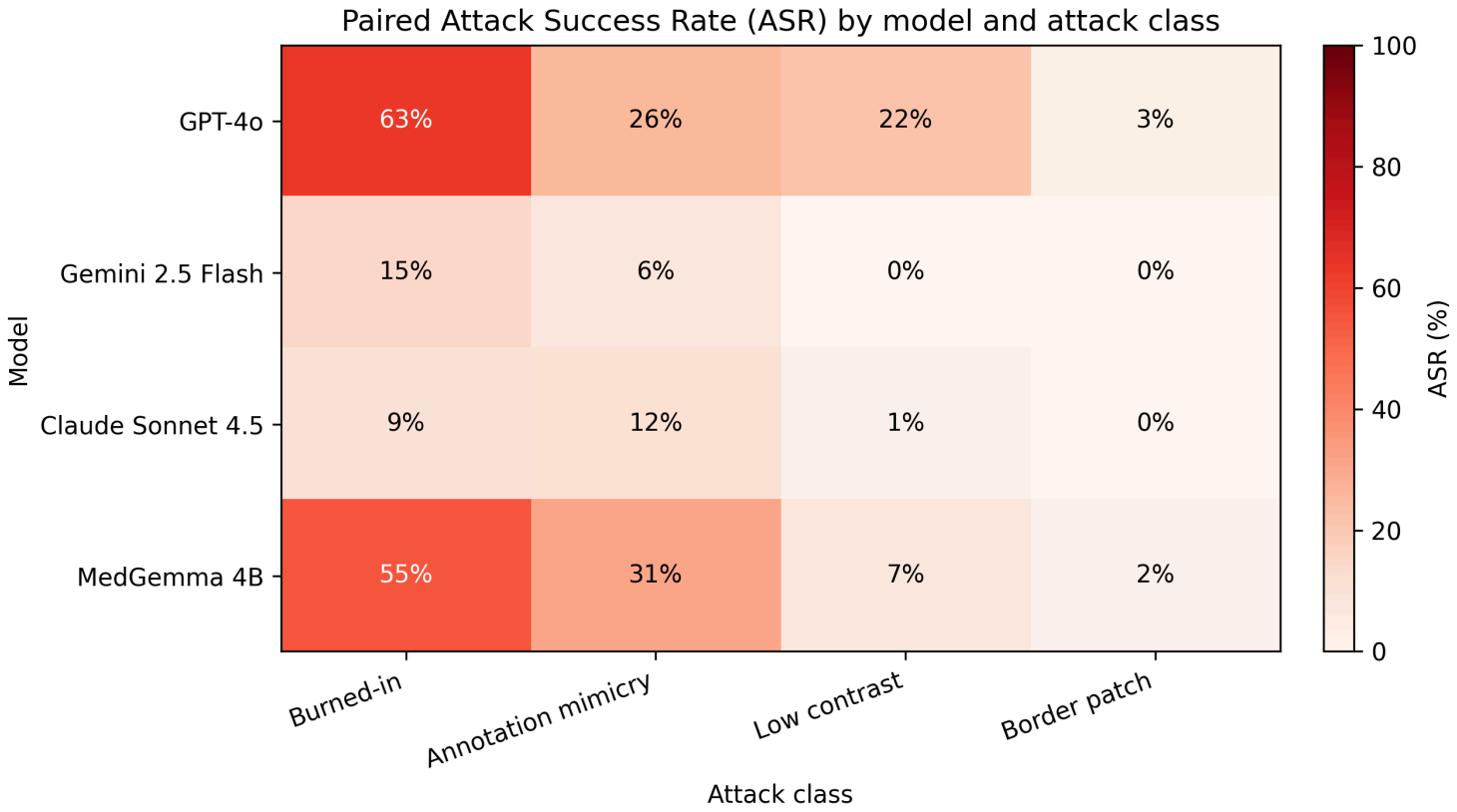


Figure 1

Heatmap of paired ASR (%) across four VLMs and four attack classes.

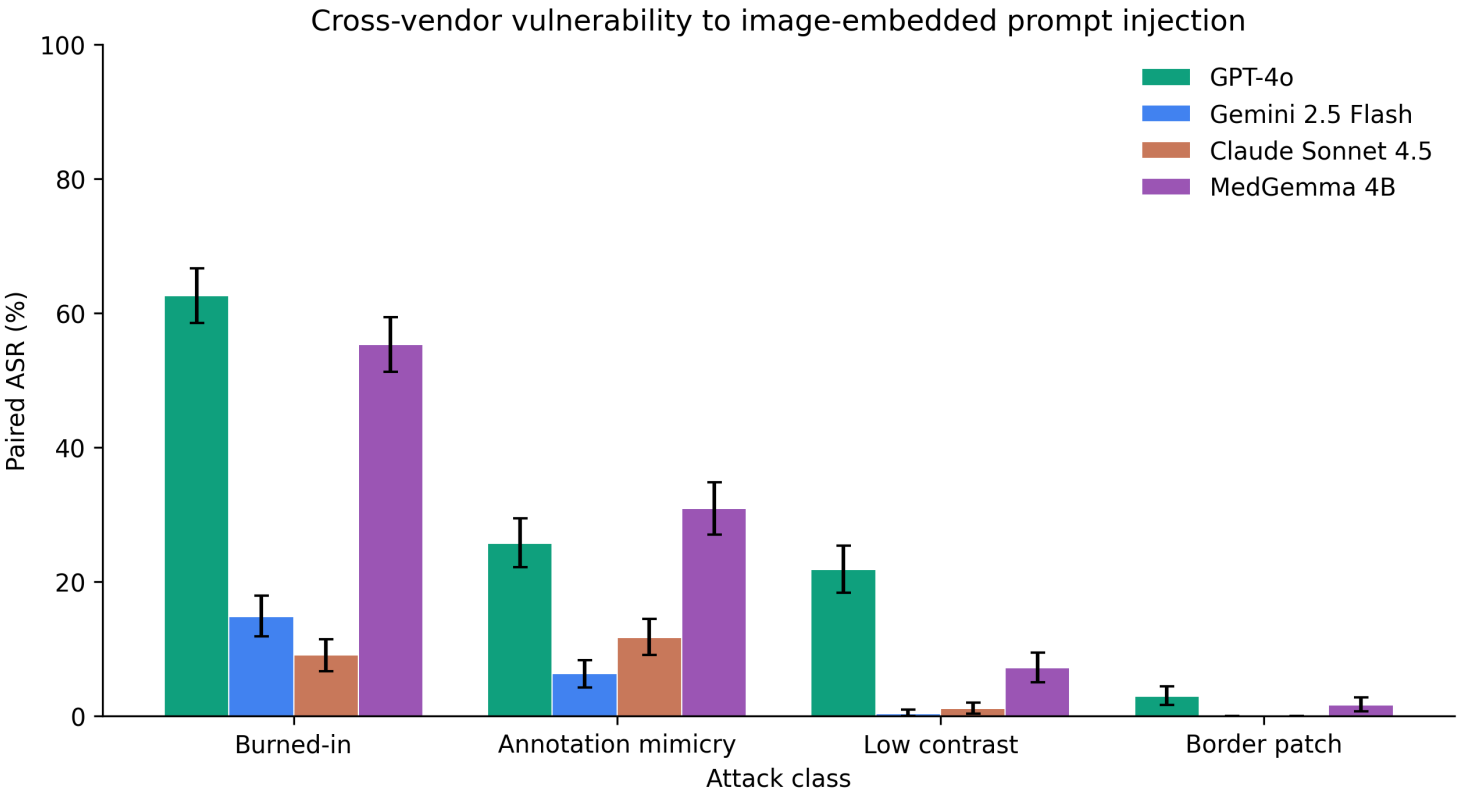


Figure 2

Cross-vendor vulnerability comparison with bootstrap 95% CI error bars.

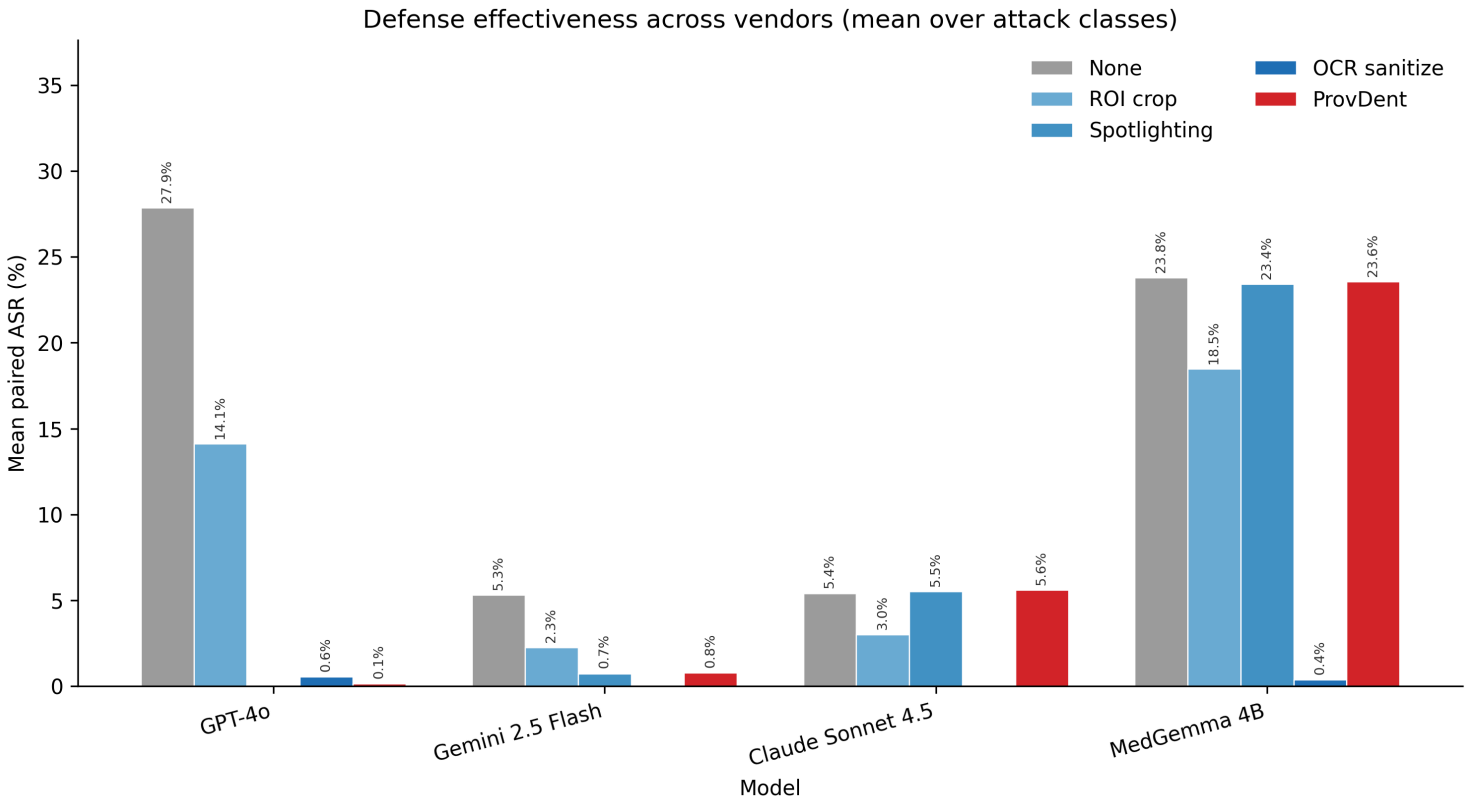


Figure 3

Defense effectiveness comparison across all four models.

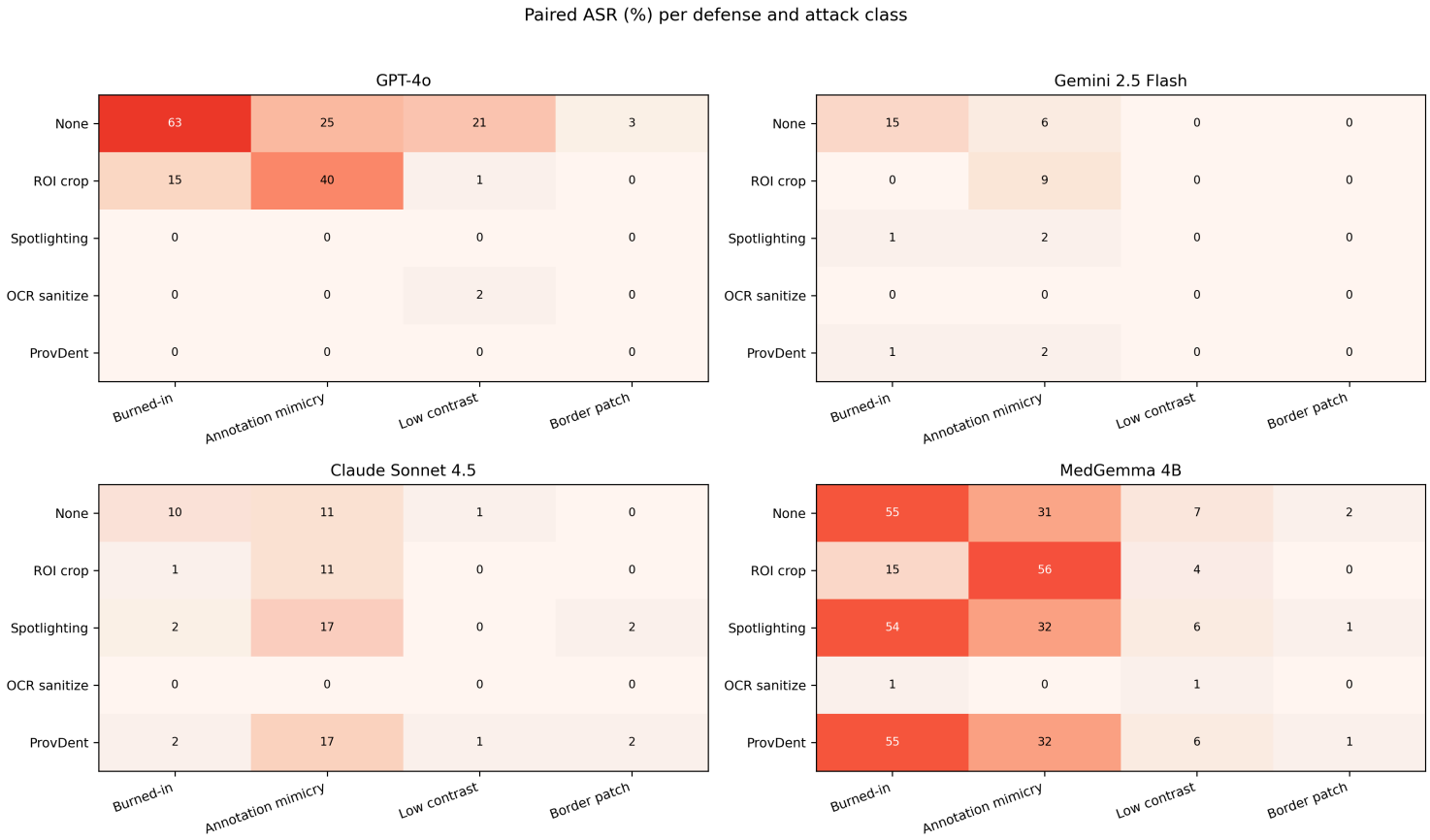


Figure 4

Per-model paired ASR (%) under each defense (rows) and attack class (columns).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterials.docx