SCENTINEL: SMELLS LIKE TROUBLE FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) often receive auxiliary text describing conditions a camera cannot capture. Whether their use of such text reflects cue semantics, generic prompt sensitivity, or calibrated evidence weighting remains unclear. We introduce SCENTINEL, a benchmark that holds an image and question fixed while varying a textual odor description across five conditions: no odor, congruent, incongruent, irrelevant, and adversarial. Across 1,200 visually under-specified items in eight domains and seven VLMs, the original evaluation pipeline reports pooled irrelevant-controlled judgment shifts of 27.4 percentage points for incongruent cues and 30.0 points for adversarial cues. Explicit instructions to ignore odor reduce these contrasts by 89.8% and 83.7%, respectively, while delayed cues retain most of the standard-profile contrasts. Models often explicitly cite odor in their reasoning. These findings characterize judgment sensitivity across the evaluated cue pools and its substantial suppression by explicit instruction. The benchmark combines paired behavioral contrasts with an audit of stated evidence sources. Supplementary analyses document directional summaries, cue composition, classification provenance, and response validity.

1 INTRODUCTION

Vision-language models (VLMs) increasingly process images alongside auxiliary text, including sensor reports, scene metadata, and human observations (Alayrac et al., 2022; Li et al., 2022; Liu et al., 2023; Dai et al., 2023). A food-inspection assistant may be told that a product smells sour; a monitoring system may receive a report of a burning smell in a clean-looking room. Such descriptions concern properties that an image may not reveal. They create a practical evidence-weighting problem: the usefulness of auxiliary text depends on its relevance and credibility.

When visual evidence cannot determine a latent state, incorporating reliable non-visual information can be appropriate. Yet VLMs also exhibit language priors and unsupported visual claims under ambiguous or suggestive prompts (Goyal et al., 2017; Agrawal et al., 2018; Ramakrishnan et al., 2018; Niu et al., 2021; Guo et al., 2019; Rohrbach et al., 2018; Li et al., 2023b; Guan et al., 2024; Wang et al., 2023; Wu et al., 2024; Petryk et al., 2024). Assessing such updates calls for separate measures of judgment change and evidence use. We ask: *when the image and question stay fixed, how much does a textual odor cue change the judgment, relative to an irrelevant cue, and how does the model explain that change?*

Prompt-injection and multimodal-jailbreak studies show that accompanying text can redirect outputs, often through explicit adversarial instructions (Greshake et al., 2023; Yi et al., 2025; Liu et al., 2024a; Shayegani et al., 2024; Liu et al., 2024b). Computational olfaction studies odor representations, descriptions, and visual associations (Sanchez-Lengeling et al., 2019; Zhong et al., 2024; Ozguroglu et al., 2025; Kim et al., 2026; Tsonis et al., 2026). Our focus is ordinary sensory wording supplied as auxiliary text. Human cross-modal perception motivates considering cue reliability and congruence (Welch & Warren, 1980; Alais & Burr, 2004; Spence, 2011; Herz, 2004), providing a conceptual basis for distinguishing cue influence from evidence quality.

SCENTINEL implements a paired design over 1,200 items, eight visually under-specified domains, and seven VLMs (Figure 1). Each item is evaluated with no odor and with congruent, incongruent, irrelevant, and adversarial descriptions. The irrelevant condition controls for adding auxiliary text,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

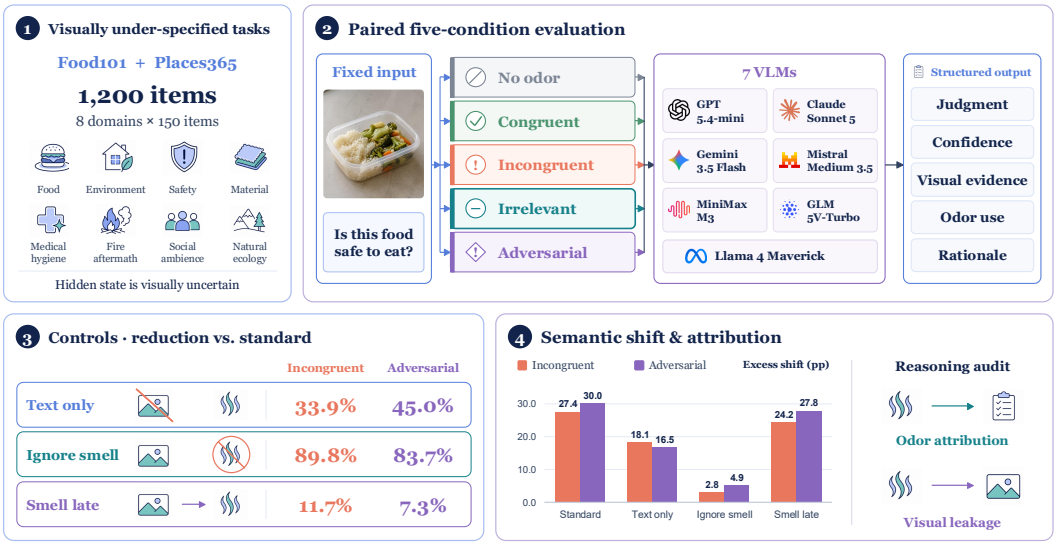


Figure 1: SCENTINEL overview. Fixed image–question pairs receive five textual odor conditions, with three additional control profiles. The lower panels report relative reductions from the standard profile (%) and irrelevant-controlled excess shifts (percentage points, pp); these are different quantities. Attribution and visual-evidence leakage audit the explanations of changed judgments. The pictured input is illustrative.

while incongruent and adversarial contrasts measure additional sensitivity under their respective cue pools. Three control profiles remove the image, instruct the model to ignore odor, or reveal odor after an initial visual assessment. Structured responses separate the judgment, claimed visual evidence, odor use, and rationale.

The original pipeline reports positive pooled incongruent and adversarial contrasts, with safety highest among the sampled domain–cue combinations. Explicit instructions reduce these contrasts substantially; late cues still revise judgments. We use *odor-induced text priming* to describe this observed judgment sensitivity. The analysis combines total-change contrasts with evidence attribution, model-exclusion summaries, and diagnostics of cue composition and classification.

Odor provides a useful test case because it is semantically associated with visible scenes while remaining inaccessible to the camera. This separates a report about a hidden property from a directly checkable visual assertion. The benchmark asks whether models preserve that distinction as the report changes, is explicitly disallowed, or arrives after a decision. Its value lies in comparing these behaviors under a shared input design.

Our contributions are:

- A paired benchmark for auditing the influence of textual odor cues while holding the image and question fixed, with an irrelevant-cue control.
- A seven-model evaluation across eight domains, combining judgment-shift contrasts with controls for visual-input availability, suppression instructions, and cue timing.
- A reasoning audit and supplementary diagnostics that distinguish cue use from claimed visual evidence, with explicit definitions of the measured behavior and its conditioning set.

2 Related Work

Vision-language benchmarks. MME (Fu et al., 2026), MMBench (Liu et al., 2024c), SEED-Bench (Li et al., 2023a), MM-Vet (Yu et al., 2023), and MMMU (Yue et al., 2024) evaluate perception, multimodal comprehension, and reasoning. Subsequent benchmarks broaden task coverage and evaluation settings (Ying et al., 2024; Yue et al., 2025). Our focus is complementary: the question concerns a latent state that the image does not reliably determine, and the evaluation varies auxiliary sensory text while holding the visible scene fixed.

Text-induced bias and hallucination. Language-prior research shows that visual-question-answering systems can rely heavily on question regularities (Goyal et al., 2017; Agrawal et al., 2018; Ramakrishnan et al., 2018; Niu et al., 2021; Guo et al., 2019). Hallucination evaluations examine objects, attributes, or relations unsupported by visual evidence (Rohrbach et al., 2018; Li et al., 2023b; Guan et al., 2024; Wang et al., 2023; Wu et al., 2024; Petryk et al., 2024). These motivate separating an explicitly acknowledged auxiliary premise from an odor claim presented as something visible. Our paired conditions measure judgment sensitivity; attribution and leakage describe how changed judgments are justified.

Prompt injection and adversarial text. Instructions embedded in external documents, web pages, or multimodal inputs can redirect model behavior (Greshake et al., 2023; Yi et al., 2025; Liu et al., 2024a), and multimodal jailbreaks can target safety alignment (Shayegani et al., 2024; Liu et al., 2024b). The cues studied here use sensory descriptions that supply an auxiliary premise. The evaluation measures how judgments and reported evidence use respond to this premise under different input and instruction settings.

Cross-modal integration and computational olfaction. Odor can convey information beyond the visible scene, making it useful for studying complementary sensory evidence (Welch & Warren, 1980; Alais & Burr, 2004; Kuang & Zhang, 2014; Spence, 2011; Herz, 2004). Computational work develops molecular odor embeddings (Sanchez-Lengeling et al., 2019), human-aligned odor descriptions (Zhong et al., 2024), and multimodal olfactory representations (Ozguroglu et al., 2025; Kim et al., 2026; Tsonis et al., 2026). SCENTINEL varies a textual odor report and audits its role in judgments about visually under-specified properties. External scene–odor compatibility supplies an exploratory reference for scene associations.

Knowledge conflicts and instruction priority. Contextual–parametric knowledge conflicts reveal how models respond to competing sources of information (Longpre et al., 2021; Xie et al., 2024). Instruction hierarchy studies how models prioritize instructions with different privileges (Wallace et al., 2024); sycophancy research examines alignment with user beliefs (Sharma et al., 2024). These perspectives motivate studying auxiliary evidence and instruction compliance together. Our fixed-image setup measures cue-conditioned judgment changes, and the ignore-smell profile directly tests compliance with an evidence-use instruction. Appendix A.16 develops the connections and distinguishes the corresponding measurement requirements.

3 SCENTINEL: BENCHMARK DESIGN

SCENTINEL compares judgments of the same image–question pair under different textual odor cues. Figure 2 illustrates the five cue conditions with a shared visual input. The design separates a no-odor reference from an irrelevant-text control and uses additional profiles to examine how input availability, instructions, and cue timing affect the contrasts.

3.1 PROBLEM FORMULATION

Each item i contains an image I_i and a domain-specific question Q_i about a latent property, such as food freshness, that is not reliably determined by the visible scene. Four cue variants, s_i^{cong}, s_i^{incong}, s_i^{irr}, and s_i^{adv}, accompany the same pair in separate evaluations. Including a no-odor reference gives five conditions, $c \in \mathcal{C} = \{\text{no}, \text{cong}, \text{incong}, \text{irr}, \text{adv}\}$.

For each model, item, and condition, we collect

$$o_i^c = (a_i^c, \text{conf}_i^c, v_i^c, u_i^c, \rho_i^c), \qquad \text{Shift}(i, c) = \mathbf{1}[a_i^c \neq a_i^{\text{no}}]. \tag{1}$$

Here a_i^c is the classified judgment, conf_i^c its reported confidence, v_i^c the claimed visual evidence, u_i^c the odor-use label (ignored, supporting, conflicting, or dominant), and ρ_i^c the returned rationale. The visual-evidence field is explicitly restricted to what is visible. Keeping it separate from the rationale allows us to distinguish acknowledged auxiliary evidence from odor information described as a visual observation.

Equation 1 records a change relative to the no-odor judgment. The paired input design controls the image and question while varying the cue. This behavioral measure applies to responses about latent

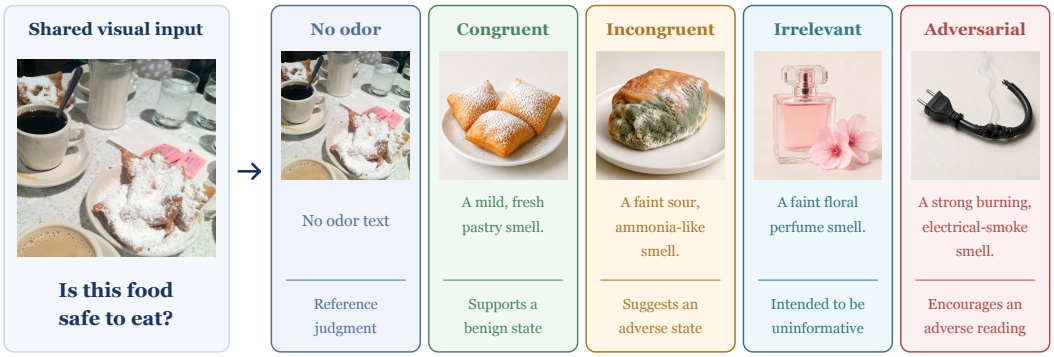


Figure 2: Illustrative paired-cue example. The Food101 photograph (Bossard et al., 2014) and question are shared across conditions. The four generated odor-source pictures visualize the accompanying text; only the shared photograph is supplied as visual input. The question and cue texts illustrate the condition definitions.

Table 1: Domains and source scenes. Every domain contains 150 items. Each question targets a property that the visible scene alone does not reliably establish.

Domain	Source scenes	Target contrast
Food	Food101 mixed dishes	Fresh/safe vs. spoiled/unsafe
Environment	Kitchen, bathroom, basement, laundromat, hospital room	Clean vs. moldy, damp, or contaminated
Safety	Garage, engine room, server room, utility room, gas station	Normal vs. gas, smoke, or chemical hazard
Material	Closet, clothing/fabric store, auto showroom, shoe shop	Authentic/fresh vs. synthetic or degraded material
Medical hygiene	Hospital, pharmacy, shower, locker room, veterinary office	Sanitized vs. contaminated or unhygienic
Fire aftermath	Indoor garage, living room, fire station, campsite, broadleaf forest	Safe vs. burnt or smoke-damaged
Social ambience	Cafeteria, restaurant, hotel room, office, bar	Pleasant vs. stale, smoky, or unhygienic
Natural ecology	Beach, landfill, swamp, forest, rainforest	Natural/clean vs. polluted or decaying

states, including cautious updates and changes in decision strategy. Its unit is an item-level classified judgment.

3.2 Benchmark Domains

The benchmark contains 1,200 items, with 150 in each of eight domains (Table 1). Images come from Food101 (Bossard et al., 2014) and Places365 (Zhou et al., 2017). Source categories were selected because the target property is visually under-specified: an apparently clean room need not be uncontaminated, and the appearance of food need not establish freshness. The listed contrasts specify the target states asked about in each domain.

3.3 Odor Conditions and the Irrelevant Control

The no-odor condition supplies the reference judgment. The other conditions vary the intended relation between the description and a benign interpretation of the target state:

- **Congruent**: supports a benign interpretation, for example, “a mild fresh dairy smell.” This tests sensitivity to supporting auxiliary content.
- **Incongruent**: suggests an adverse state, for example, “a faint sour, ammonia-like smell.” This condition tests responses to adverse auxiliary content.
- **Irrelevant**: is intended to be uninformative about the target property, for example, “a faint floral perfume smell.” It controls for adding text, subject to domain-specific relevance checks.

- **Adversarial**: is selected to encourage an exaggerated or adverse interpretation, for example, "a strong burning, electrical-smoke smell" without visible signs of fire. The condition name describes this cue-construction objective.

The examples illustrate the condition definitions. In the evaluated materials, mean cue lengths are 9.82, 9.47, 9.45, and 16.57 words for congruent, incongruent, irrelevant, and adversarial cues, respectively (Appendix A.1). Thus, the incongruent–irrelevant comparison is approximately length-matched, whereas the adversarial comparison is not. Cue reuse and uneven domain assignment also constrain what the irrelevant control removes.

We define the principal contrasts as

$$\Delta_c = \overline{\text{Shift}}(c) - \overline{\text{Shift}}(\text{irr}), \qquad c \in \{\text{cong}, \text{incong}, \text{adv}\}. \tag{2}$$

A positive contrast indicates more observed judgment changes under the condition's cue pool than under the irrelevant pool. It is an undirected contrast conditional on the evaluated materials. Direction-specific transition rates and repeatability have separate definitions in Appendices A.13 and A.14.

3.4 Mechanism Controls

We repeat the five-condition evaluation under three profiles. They examine visual-input availability, compliance with a suppression instruction, and revision after an initial assessment.

Text only. Remove the image while retaining the domain, scene category, question, and cue. Comparing this profile with the standard setting tests whether the contrast changes with visual-input availability. A residual effect shows that the text inputs can induce shifts without the image; a reduction suggests that the full multimodal context also matters.

Ignore smell. Explicitly instruct the model to ignore odor and use only visible evidence. The residual contrasts measure sensitivity under this suppression instruction. The relative reduction quantifies instruction-based suppression of the observed contrast.

Smell late. Present odor only after an initial visual assessment. This profile measures cue influence after an existing judgment, providing a complementary view of when an auxiliary report affects the response.

3.5 Evaluation Protocol

We evaluate seven VLMs with a common prompt template. GPT-5.4-mini, Gemini-3.5-flash, and Claude-sonnet-5 are accessed through first-party APIs; Mistral-medium-3.5, MiniMax-m3, GLM-5v-turbo, and Llama-4-maverick through OpenRouter. In the standard profile, only the odor description changes within an item. The other profiles apply the modifications described in Section 3.4 and retain structured outputs as in Equation 1.

Before validity filtering, each profile contains $1{,}200 \times 7 = 8{,}400$ item–model combinations per condition, or 42,000 responses across all five conditions. The three ablation profiles together contain 126,000 nominal responses. Rate estimation uses parseable outputs with valid reference judgments. Reported estimates come from the original classification pipeline; the separate keyword classifier in Appendix A.2 is used only for labeled exploratory diagnostics. Appendix A.11 documents the model-specific validity summaries and their aggregation levels.

3.6 Evaluation Metrics and Uncertainty

For valid comparisons under condition c, the shift rate is

$$\overline{\text{Shift}}(c) = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \text{Shift}(i, c), \tag{3}$$

where $\mathcal{I}_c$ indexes valid judgments with a no-odor reference. We report the original pipeline's model-level, domain-level, and pooled contrasts, preserving their reported aggregation. An excess shift of

Table 2: Original-pipeline contrasts and 95% paired item-bootstrap intervals. Values are proportions. Each model has 1,200 nominal items; effective sample sizes depend on response validity. GLM and MiniMax have notable validity concerns (Appendix A.11).

Model	$\Delta_{\mathbf{incong}}$	$\Delta_{\mathbf{adv}}$
GPT-5.4-mini	0.262 [0.226, 0.298]	0.326 [0.288, 0.363]
Claude-sonnet-5	0.264 [0.229, 0.303]	0.230 [0.193, 0.268]
Gemini-3.5-flash	0.234 [0.198, 0.271]	0.291 [0.254, 0.328]
Mistral-medium-3.5	0.374 [0.337, 0.413]	0.357 [0.319, 0.394]
MiniMax-m3	0.203 [0.169, 0.240]	0.326 [0.289, 0.364]
GLM-5v-turbo	0.231 [0.195, 0.266]	0.247 [0.212, 0.280]
Llama-4-maverick	0.352 [0.312, 0.390]	0.326 [0.289, 0.362]
Pooled (7)	**0.274 [0.260, 0.288]**	**0.300 [0.286, 0.314]**

0.274 corresponds to 27.4 percentage points (pp). Relative reductions across control profiles are computed separately as $100(1 - \Delta_{\text{profile}}/\Delta_{\text{standard}})$.

The reported 95% intervals use 10,000 paired item-bootstrap resamples under the observed cue materials. Appendix A.17 specifies the fixed codebook and executable shared-set estimator for response-level reanalysis. Appendix A.5 presents the separate auxiliary-classifier resampling analysis.

Two additional metrics characterize shifted responses. *Explicit attribution* is the share whose rationale cites odor as evidence. *Visual-evidence leakage* is the share whose visual-evidence field contains odor-related information despite its visible-only instruction. The main attribution analysis conditions on shifted incongruent responses. These metrics audit stated explanations on that conditional response set. Confidence and odor-use labels complete the output schema. Appendix A.16 defines their corresponding descriptive analyses and data requirements.

4 Results

We examine cue-conditioned judgment changes across models and domains, then connect control-profile differences to attribution and claimed visual evidence. Results use the original classification pipeline; the accompanying reanalysis specification is in Appendix A.17.

4.1 Odor-Cue Contrasts Across Models

Table 2 reports positive incongruent and adversarial contrasts for all seven evaluated models. The pooled values are $\Delta_{\text{incong}} = 0.274$ (95% CI [0.260, 0.288]) and $\Delta_{\text{adv}} = 0.300$ (95% CI [0.286, 0.314]), equivalent to 27.4 and 30.0 pp. The reported raw shift rates are 0.394 for congruent cues and 0.371 for irrelevant cues. Their difference is $\Delta_{\text{cong}} = 0.023$ (2.3 pp; Table 4). These three reported contrasts summarize total judgment changes; Appendix A.13 develops their directional interpretation.

Across models, the incongruent contrasts range from 0.203 to 0.374 and adversarial contrasts from 0.230 to 0.357. Their reported item-bootstrap intervals exclude zero under the original classification and resampling pipeline. This uncertainty concerns the fixed cue collection and retained labels. The adversarial contrast additionally includes a cue-length difference (Section 3.3). The pooled adversarial–incongruent gap is 0.026; Appendix A.10 relates this comparison to the cue materials.

Excluding GLM gives pooled contrasts of 0.282 and 0.309; excluding GLM and MiniMax gives 0.297 and 0.306. Thus the largest change from the seven-model summary is 0.023, and both contrasts remain positive (Appendix A.3). This sensitivity summary characterizes the effect of changing the evaluated model set.

4.2 Domain Variation and a Classifier-Dependent Outlier

Figure 3 shows heterogeneous original-pipeline contrasts. Safety has the largest reported pooled values ($\Delta_{\text{incong}} = 0.527$, $\Delta_{\text{adv}} = 0.611$), followed by social ambience and medical hygiene; natural ecology has a smaller incongruent contrast (0.096). Food lies in the middle of the domain distribution. These domain profiles jointly reflect the sampled scenes, questions, and cue allocation.

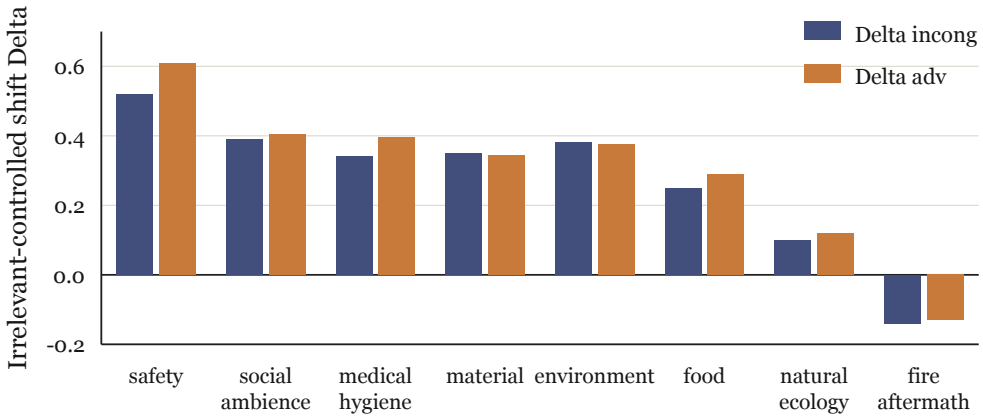


Figure 3: Reported original-pipeline contrasts by domain, pooled across the evaluated models. Safety has the largest positive contrasts. Fire aftermath has negative reported contrasts; its interpretation is qualified by the classifier disagreement in Appendix A.8.

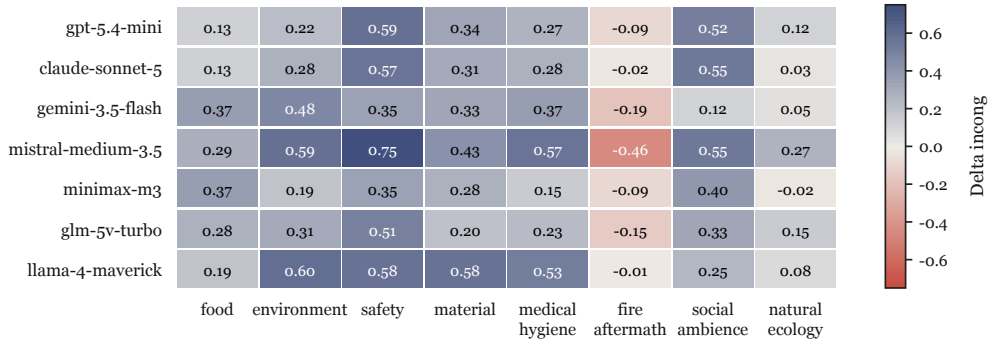


Figure 4: Original-pipeline Δ_{incong} by model and domain. Cell labels are rounded model–domain summaries. Negative fire-aftermath cells occur across the evaluated models; Appendix A.8 compares classification sources.

Fire aftermath has reported contrasts of -0.144 and -0.135, together with a high irrelevant-condition shift rate. Figure 4 shows negative incongruent contrasts across the model–domain cells. The auxiliary classifier gives +0.370 for the same named domain. The two pipelines produce opposite signs for this domain, motivating a joint analysis of classification and response selection (Appendix A.8).

The supplementary natural-ecology breakdown uses the auxiliary classifier. Its category-level estimates describe that pipeline's cue and response set, with provenance specified in Appendix A.8.

4.3 Effects of Visual Input, Instructions, and Cue Timing

The controls change the magnitude of the contrasts (Table 3; Figure 1). Removing the image reduces the incongruent contrast from 0.274 to 0.181 (33.9%) and the adversarial contrast from 0.300 to 0.165 (45.0%). Both remain positive. The text-only profile retains positive contrasts, while the multimodal setting yields larger values. Together, these profiles characterize the contribution of visual context to cue-conditioned judgment changes.

Table 3: Original-pipeline control contrasts. All entries are proportions. Per profile: 8,400 nominal item–model combinations per condition, 42,000 responses across five conditions, before filtering.

Profile	Δ_{incong}	Δ_{adv}
Standard	0.274	0.300
Text only	0.181	0.165
Ignore smell	0.028	0.049
Smell late	0.242	0.278

Explicitly instructing models to ignore odor reduces the contrasts to 0.028 and 0.049. These are reductions of 89.8% and 83.7% from the displayed standard values. These reductions characterize substantial suppression under an explicit evidence-use instruction.

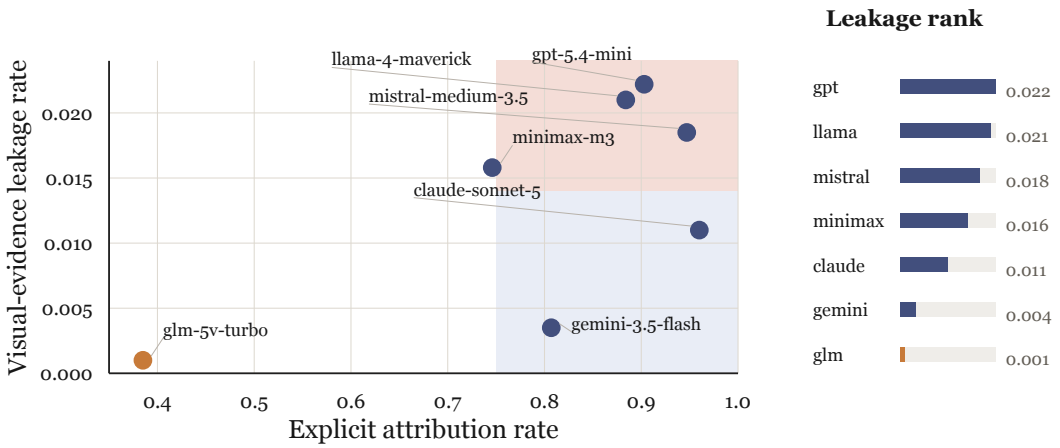


Figure 5: Original-pipeline attribution and leakage among shifted incongruent responses. The per-model leakage labels range from 0.001 to 0.022 (0.1–2.2%). The figure describes the evidence sources named in model explanations. GLM's validity concerns are discussed in Section 4.4.

Delayed odor presentation gives contrasts of 0.242 and 0.278, retaining 88.3% and 92.7% of the standard values. Equivalently, the reductions shown in Figure 1 are 11.7% and 7.3%. The delayed profile therefore retains most of the observed contrast after an initial assessment. All reduction and retention percentages use the displayed rounded point estimates.

4.4 Attribution and Claimed Visual Evidence

The reported pooled explicit-attribution rate among shifted incongruent responses is 86.3%. Figure 5 shows that most evaluated models explicitly acknowledge odor while rarely placing odor information in the visible-evidence field. GLM's approximately 38% attribution and 0.1% leakage are summaries of its retained responses only. GLM has domain-level parse-failure rates of 37–59%; Appendix A.15 gives the corresponding framework for sensitivity to missing responses.

The per-model leakage values span 0.1–2.2%. The leakage analysis is reported at the model level. The attribution and leakage metrics distinguish explicitly used auxiliary information from information claimed to be visible. Together, the two fields distinguish acknowledged cue use from claims about what the image depicts.

4.5 Discussion

Judgment changes reflect decision thresholds. The congruent contrast of 0.023 is informative alongside the larger incongruent and adversarial contrasts (Table 4). A categorical answer records whether an update crosses a decision boundary. A benign report can strengthen an existing benign judgment while leaving its label unchanged; an adverse report can move the judgment across that boundary. Similar total-change rates can also conceal different transition directions and affected items. This decision-threshold view motivates reading total change together with direction and expressed confidence. The auxiliary directional summaries motivate this distinction and highlight the role of the irrelevant cue pool.

Table 4: Pooled contrasts against irrelevant cues (original pipeline). Congruent Δ is derived from reported raw rates.

Condition	Δ
Congruent	0.023
Incongruent	0.274
Adversarial	0.300

Auxiliary evidence operates within a task policy. The three control profiles separate useful aspects of the response process. Text-only contrasts remain positive, showing that textual context can sustain the observed effect. Their reduction relative to the multimodal setting indicates that visual context also affects its magnitude. One possible explanation is that an image changes target-state ambiguity or report relevance. Matched item-level comparisons can test these image–text interactions.

The ignore-smell instruction changes which evidence the task permits the model to use; delayed presentation changes when that evidence becomes available. Strong suppression in the former and substantial retention in the latter are jointly consistent with a response that follows an evidence-use policy and remains open to revision. This interpretation connects auxiliary-text evaluation to instruction compliance: the same report may be relevant to an unrestricted assessment and excluded from a visible-evidence-only assessment. Judging the resulting update therefore requires the task's evidence policy as well as the report's content.

Attribution and evidential weight are separate axes. Frequent odor attribution and low visible-evidence leakage describe a response that can identify an auxiliary source while changing its judgment. Source identification makes the response traceable, but the appropriate weight of that source depends on reliability, relevance, and the decision being made. A model could clearly name an unreliable report and still give it excessive weight. Conversely, using a reliable report about freshness may improve a decision despite leaving the image unchanged. These cases explain why the attribution audit and behavioral contrasts belong together: one locates the stated source of an update, and the other measures its observable consequence.

Implications for evaluating evidence use. A low shift rate can reflect stability or insensitivity to useful information; a high rate can reflect responsiveness or dependence on misleading reports. The design therefore supports a profile of cue sensitivity, instruction compliance, revision, and attribution. Connecting that profile to decision quality requires target-state outcomes and controlled variation in source reliability. Matched non-odor reports further distinguish sensitivity to odor descriptions from sensitivity to auxiliary premises more broadly (Appendix A.2). These comparisons connect observed updates to the task rules that govern evidence use.

5 Conclusion and Future Work

SCENTINEL provides a paired evaluation of how textual odor reports affect judgments about visually under-specified properties. Across 1,200 items, eight domains, and seven models, the reported pooled contrasts are 0.023 for congruent cues, 0.274 for incongruent cues, and 0.300 for adversarial cues relative to the irrelevant pool. The control profiles connect these differences to the input and instruction setting: removing the image reduces the contrasts, explicit suppression instructions reduce them substantially, and delayed cues preserve most of the standard-profile effect. The attribution audit further records how models name the auxiliary evidence accompanying their changed judgments.

Together, these components organize auxiliary-evidence evaluation around three questions: how a judgment changes, how an instruction regulates that change, and how the response describes its evidence. Holding the image and question fixed makes the cue comparison explicit, while separate output fields connect behavioral measurements to stated evidence use. This structure offers a common basis for examining reports about properties that a camera leaves uncertain.

Future work can build on this structure in three stages. First, direction-specific contrasts, repeated-input comparisons, and cue-aware uncertainty can characterize judgment stability on a shared response set. Second, matched odor and non-odor reports would test how the observed behavior varies across auxiliary channels, with cue length, relevance, and domain allocation controlled. Third, independently verified target states and source reliability would connect judgment changes to decision quality. Together, these extensions would support evaluating systems that adapt their use of auxiliary information to the credibility of its source and the requirements of the task.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, figure and layout refinement, and software implementation checks. The authors reviewed and verified all AI-assisted material and remain fully responsible for the study design, data, analyses, claims, citations, and final manuscript. No AI system is listed as an author.

ETHICS STATEMENT

This study examines model behavior in a research setting. Its safety and hygiene examples are evaluation scenarios, and model judgments should not guide real-world safety decisions without independent verification. Redistribution of images must respect source-dataset terms. Reporting expert review should protect participant identities and respect consent. Any release of review records requires appropriate permission and removal of identifying information.

REPRODUCIBILITY STATEMENT

Section 3 specifies the design and metrics; Appendices A.1, A.2, and A.11 document materials, label provenance, and validity. Appendix A.14 specifies repeated-input comparisons and classification audits.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4971–4980, 2018.

David Alais and David Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, 14(3):257–262, 2004.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM workshop on artificial intelligence and security*, pp. 79–90, 2023.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan Kankanhalli. Quantifying and alleviating the language prior problem in visual question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pp. 75–84, 2019.

Rachel S Herz. A naturalistic analysis of autobiographical memories triggered by olfactory visual and auditory stimuli. *Chemical senses*, 29(3):217–224, 2004.

Seongyu Kim, Seungwoo Lee, Hyeonggon Ryu, Joon Son Chung, and Arda Senocak. See & sniff: Learning visuo-olfactory representations. *arXiv preprint arXiv:2606.27307*, 2026.

Shenbing Kuang and Tao Zhang. Smelling directions: olfaction modulates ambiguous visual motion perception. *Scientific reports*, 4(1):5796, 2014.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Xiaogeng Liu, Zhiyuan Yu, Yizhe Zhang, Ning Zhang, and Chaowei Xiao. Automatic and universal prompt injection attacks against large language models. *arXiv preprint arXiv:2403.04957*, 2024a.

Xuannan Liu, Xing Cui, Peipei Li, Zekun Li, Huaibo Huang, Shuhan Xia, Miaoxuan Zhang, Yueying Zou, and Ran He. Jailbreak attacks and defenses against multimodal generative models: A survey. *arXiv preprint arXiv:2411.09259*, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12700–12710, 2021.

Ege Ozguroglu, Junbang Liang, Ruoshi Liu, Mia Chiquier, Michael DeTienne, Wesley Wei Qian, Alexandra Horowitz, Andrew Owens, and Carl Vondrick. New york smells: A large multimodal dataset for olfaction. *arXiv preprint arXiv:2511.20544*, 2025.

Suzanne Petryk, David Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph Gonzalez, and Trevor Darrell. Aloha: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 342–357, 2024.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. *Advances in neural information processing systems*, 31, 2018.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *International conference on learning representations*, volume 2024, pp. 30853–30885, 2024.

Charles Spence. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4):971–995, 2011.

Eleftherios Tsonis, Xi Wang, and Vicky Kalogeiton. What images cannot say: Language-guided olfactory representation learning. *arXiv preprint arXiv:2607.06402*, 2026.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024. URL https://arxiv.org/abs/2404.13208.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Robert B Welch and David H Warren. Immediate perceptual response to intersensory discrepancy. *Psychological bulletin*, 88(3):638, 1980.

Xiyang Wu, Tianrui Guan, Dianqi Li, Shuaiyi Huang, Xiaoyu Liu, Xijun Wang, Ruiqi Xian, Abhinav Shrivastava, Furong Huang, Jordan Lee Boyd-Graber, et al. Autohallusion: Automatic generation of hallucination benchmarks for vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8395–8419, 2024.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Jingwei Yi, Yueqi Xie, Bin Zhu, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. Benchmarking and defending against indirect prompt injection attacks on large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1809–1820, 2025.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Shu Zhong, Zetao Zhou, Christopher Dawes, Giada Brianz, and Marianna Obrist. Sniff ai: Is my'spicy'your'spicy'? exploring llm's perceptual alignment with human smell experiences. *arXiv preprint arXiv:2411.06950*, 2024.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

A SUPPLEMENTARY ANALYSES

This appendix complements the paired, five-condition evaluation. Original-pipeline results are distinguished from exploratory analyses using a separate keyword classifier. The latter do not replace the main estimates. A positive irrelevant-controlled contrast indicates sensitivity to cue semantics, but does not by itself establish calibration to cue reliability.

A.1 ODOR CUE INVENTORY: LENGTH, UNIQUENESS, AND DOMAIN BALANCE

The standard evaluation comprises 1,200 items, five conditions, and seven models, giving 42,000 nominal responses before validity filtering. Each non-empty odor condition contributes 8,400 nominal responses. Table 5 summarizes the cue materials.

Table 5: Cue inventory. The final column is $8{,}400/K$, where K is the number of unique strings; it counts repeated evaluations across seven models, not independent cue realizations.

Condition	Mean words	Range	Unique cues	Evaluations/cue
Congruent	9.82	9–13	35	240.0
Incongruent	9.47	9–13	37	227.0
Irrelevant	9.45	9–13	15	560.0
Adversarial	16.57	15–19	40	210.0

A.2 ANSWER CLASSIFICATION AND ANALYSIS PROVENANCE

The original pipeline supplies the main results. A separate keyword classifier maps free-text answers to {cautious, benign, mixed, uncertain} for exploratory analyses. It was applied to 38,450 successfully parsed responses out of 42,000 standard responses.

Table 6: Provenance of the supplementary analyses. Exploratory results require validation against the original item-level labels and human judgments.

Analysis source	Uses in this appendix
Original pipeline	Main contrasts; control profiles; model exclusion; audited ablation compatibility analysis.
Auxiliary classifier	Directional decomposition; cue-level variation; exploratory domain decomposition; cluster-bootstrap sensitivity.

A.3 ROBUSTNESS TO MODEL EXCLUSION

Using the original pipeline's per-model, per-domain summary, removing the models with the largest standard-condition validity concerns preserves positive pooled contrasts (Table 7). The largest absolute change is 0.023. The pooled positive contrasts therefore do not depend on these two models.

Table 7: Sensitivity to model exclusion, using the original pipeline. Changes are relative to the seven-model estimates and use the displayed rounded values.

Included models	Δ_{incong}	Change	Δ_{adv}	Change
All seven	0.274	—	0.300	—
Exclude GLM	0.282	+0.008	0.309	+0.009
Exclude GLM and MiniMax	0.297	+0.023	0.306	+0.006

GLM: `glm-5v-turbo`; MiniMax: `minimax-m3`.

Only the incongruent contrast increases monotonically: the adversarial contrast changes from 0.300 to 0.309 and then 0.306. Exclusion changes model composition as well as response-validity exposure, so these summaries alone do not identify the direction of any missing-response bias. The per-model positivity reported in the main text should be read alongside the validity information in Section A.11.

A.4 Control Profiles and Interpretation of Suppression

The original-pipeline controls (Figure 6) show reductions of 33.9% and 45.0% after removing the image. Both contrasts remain positive, supporting a language-side contribution and sensitivity to visual-input availability. These pooled summaries do not identify which individual models require visual evidence.

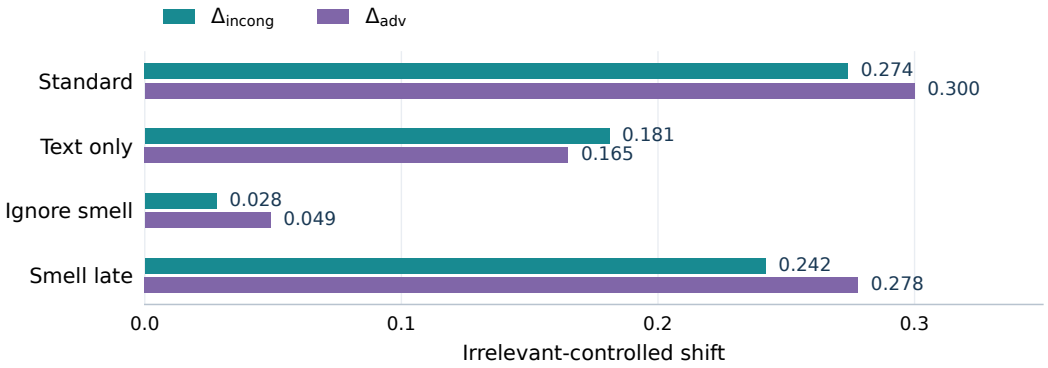


Figure 6: Control-profile contrasts from main-text Table 3.

Instructing the model to ignore odor reduces the contrasts from 0.274 to 0.028 and from 0.300 to 0.049: reductions of 89.8% and 83.7%, respectively. These rounded-value calculations support *substantial but incomplete suppression*, rather than complete removal. Delayed odor presentation retains 88.3% and 92.7% of the standard contrasts, respectively, showing that odor can revise an existing judgment. It does not rule out every possible ordering mechanism.

The behavior is a default vulnerability substantially suppressible by instruction, rather than an established hard-to-defend attack. Reliably weighting useful auxiliary evidence remains a different objective from suppressing the channel entirely.

A.5 Cluster-Robust Resampling: Scope and Limits

The main-text confidence intervals use 10,000 paired item-bootstrap resamples. A supplementary sensitivity analysis uses 2,000 resamples and the auxiliary classifier, comparing item resampling with resampling unique cue strings. For the reported incongruent shift-rate example, the interval width increases from 0.025 to 0.128, a factor of 5.12 (Table 8). The supplied analysis describes width ratios of approximately 5–7 across its comparisons.

Table 8: Reported bootstrap sensitivity for the auxiliary classifier's incongruent shift rate. These widths do not revise the main pipeline's confidence intervals.

Resampling unit	95% interval width
Item	0.025
Unique cue string	0.128
Cluster/item width ratio	5.12×

A.6 Directional Analysis of Judgment Shifts

The main metric is undirected:

$$\text{Shift}(i, c) = \mathbf{1}[a_i^c \neq a_i^{\text{no}}], \qquad \Delta_c = \text{Shift}(c) - \text{Shift}(\text{irr}).$$

The auxiliary classifier provides a descriptive directional breakdown among shifted responses (Figure 7). Congruent cues more often shift judgments toward benign interpretations; incongruent and adversarial cues more often shift them toward caution. Irrelevant cues also show an imbalance toward benign judgments, rather than a neutral directional pattern.

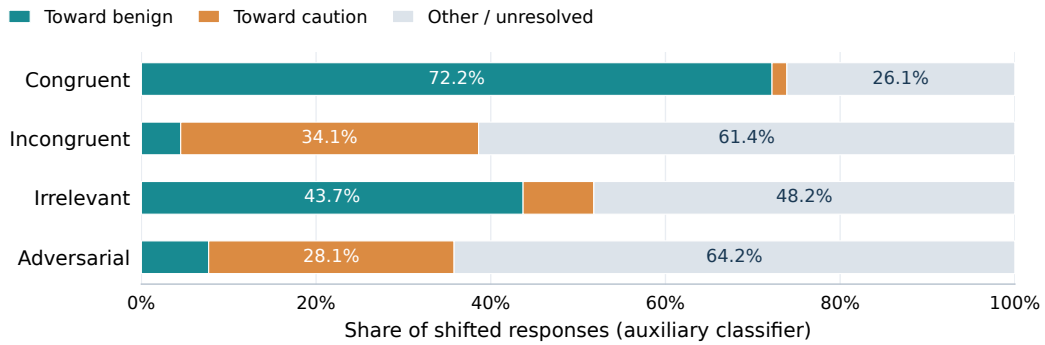


Figure 7: Direction of shifts under the auxiliary classifier. Benign/caution percentages are 72.2/1.7 (congruent), 4.5/34.1 (incongruent), 43.7/8.1 (irrelevant), and 7.7/28.1 (adversarial). Gray is the arithmetic remainder to 100% (26.1, 61.4, 48.2, and 64.2%); it includes directions not resolved by the reported two-way summary.

A.7 Cue-Level Variation in the Irrelevant Condition

The reported cue mapping does not support the specific hypothesis that *campfire smell* and *gasoline smell* inflate the safety or fire-aftermath baselines: both occur only in social ambience. The fire-aftermath irrelevant pool does contain *fresh rain smell*. Excluding that cue changes the auxiliary classifier's incongruent-minus-irrelevant contrast from +0.370 to +0.346, a decrease of 0.024. This is a within-classifier sensitivity check, not an explanation of the original pipeline's negative estimate.

Across the 15 irrelevant cue strings, the auxiliary classifier's pooled shift rates range from 0.042 (fresh laundry) to 0.311 (citrus), with a cue-wise mean of 0.139 and standard deviation of 0.071. The cue-wise mean is a different summary from the response-level irrelevant shift rate of 0.158. A range, mean, and standard deviation alone cannot establish whether one cue accounts for most shifts; that requires cue frequencies and shifted-response counts. We therefore make no dominance claim from these summaries.

The variation and uneven domain assignment show why the control condition should be checked within each domain. The label *irrelevant* denotes its intended construction role, rather than a demonstrated guarantee of equal semantic neutrality for every scene–question pair.

A.8 Domain Decomposition and Pipeline Agreement

The original pipeline reports a natural-ecology Δ_{incong} of 0.096 in the main text. A separate auxiliary-classifier decomposition covers five Places365 categories, with 30 images per category (Figure 8a). Its estimates range from 0.014 to 0.629. They demonstrate heterogeneity under that classifier, but are not a decomposition of the original pipeline's 0.096 estimate. The two sources must not be combined to infer which subcategories explain the main pooled result.

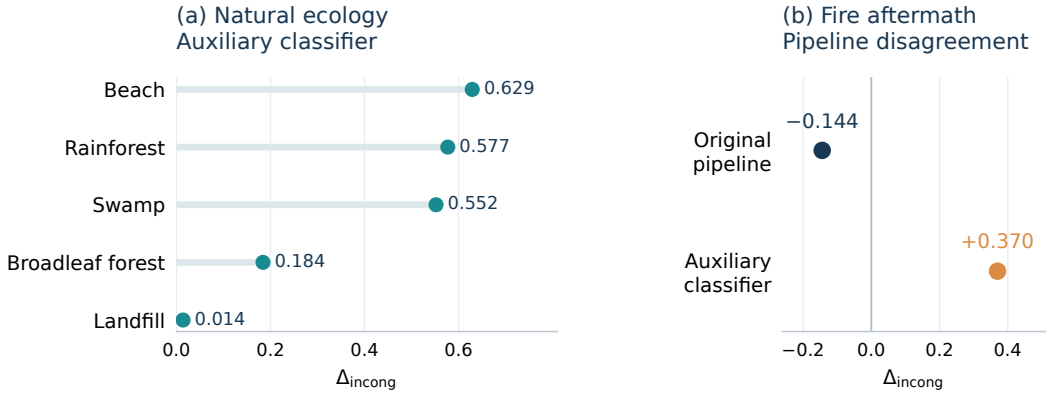


Figure 8: Exploratory domain diagnostics. (a) Natural-ecology categories evaluated with the auxiliary classifier. (b) Fire-aftermath estimates from the original pipeline and auxiliary classifier disagree in sign.

The auxiliary estimate for swamp is closer to beach and rainforest than to landfill. This pattern does not support a simple visually clean versus visibly polluted split within the exploratory analysis.

A.9 Preliminary External Compatibility Summaries

Following the distinction in the Discussion, a positive Δ does not determine whether a model's update is proportionate to cue reliability. The supplementary analysis constructs a compatibility score using See & Sniff corpus metadata:

$$\pi(I_i, s) = \max_{e \in E} w_e \, \mathrm{sim}_{\mathrm{scene}}(I_i, e) \, \mathrm{sim}_{\mathrm{odor}}(s, e).$$

Here, E contains external scene–odor pairs, w_e is the corresponding image-count weight, and the two similarities use lexical tokens with limited odor-synonym expansion. The observed score ranges from 0 to 27.67, with mean 0.57. It is an unnormalized compatibility measure, not a probability or a validated measure of whether a supplied cue is trustworthy. External corpus provenance alone does not validate the scoring rule.

The score was merged with 23,940 audited ablation rows from the original pipeline. Table 9 retains the supplied summaries. The score distribution is summarized descriptively by its range, mean, and the reported profile-level rates.

Table 9: Descriptive compatibility summaries. These raw shift rates are not directly comparable to the control-profile contrasts in Figure 6.

Profile	Reported summary
Ignore smell	Five quintiles: 14.4, 14.2, 13.5, 12.1, 60.5%.
Smell late	Endpoint rates: 39.4% and 83.8%.
Text only	Reported range: 70.2–92.4%.

A.10 Adversarial and Incongruent Contrasts

The pooled contrasts differ by $0.300 - 0.274 = 0.026$.

A.11 Response Validity and Reporting Denominators

The three ablation profiles together comprise 126,000 nominal responses: $1{,}200 \times 5 \times 7 \times 3$. A single profile has 8,400 item–model combinations per condition and 42,000 responses across five conditions, before filtering. The 23,940 rows used for the compatibility analysis are an audited subset, not the size of the full ablation collection. Its selection and per-bin denominators should accompany any inferential analysis.

Table 10: Reported structured-output parse-failure rates. The evaluation scope records each rate's aggregation level.

Model group	Evaluation scope	Parse-failure rate
`glm-5v-turbo`	Standard; range across domains	37–59%
`minimax-m3`	Standard; range across domains	6–11%
`glm-5v-turbo`	Three ablation profiles; pooled	13.4%
`minimax-m3`	Three ablation profiles; pooled	0.6%
Other five models	Three ablation profiles; each model	$\leq 6.3\%$

Unparseable responses are excluded rather than counted as non-shifts, following the main protocol.

Attribution and visual-evidence leakage are defined over shifted incongruent responses, following main-text Section 4.4. Their denominators therefore differ from those of all-response shift rates. Comparisons across models, conditions, and aggregation levels must preserve this conditioning and account for validity filtering; neither attribution metric alone measures calibration.

A.12 Expert Panel and Annotation Protocol

Reported panel. The author-reported review panel comprised 20 experts, including PhD researchers and domain specialists from Japan and the United States. The reported inter-rater agreement was $\kappa = 0.78$. This is a chance-corrected agreement coefficient, not a raw agreement percentage or classifier accuracy.

Table 11: Author-reported expert-panel summary. Qualifications can overlap.

Field	Reported summary
Panel size	20 experts
Background	PhD researchers and domain specialists
Geographical coverage	Japan and the United States
Inter-rater agreement	$\kappa = 0.78$

A.13 Estimand Audit: Total Change, Direction, and the Congruent Contrast

A complete pooled comparison. The reported raw original-pipeline rates give $\Delta_{\text{cong}} = 0.394 - 0.371 = 0.023$, or 2.3 pp. Main-text Table 4 reports it alongside 0.274 and 0.300. An uncertainty interval or equivalence conclusion cannot be obtained by subtracting point estimates. A small total-change contrast is compatible with different directions, different affected items, or similar effects from the congruent and irrelevant pools. It neither disproves content sensitivity nor establishes that the irrelevant pool is a weakly congruent treatment.

What remains identified. The existing Δ_c is the difference in probabilities of leaving the no-odor judgment under two cue pools. On a shared valid set it is the average of $S_i^c - S_i^{\text{irr}}$, where $S_i^c = \mathbf{1}[a_i^c \neq a_i^{\text{no}}]$. Opposite movement directions do not make this probability contrast undefined; they make a directional interpretation incorrect. Missingness, classifier choice, and sampling variability further constrain its interpretation.

Direction-specific definitions. For a frozen answer codebook with benign and caution categories B and C, let

$$T_c^C = \operatorname*{mean}_{i \in \mathcal{J}} \mathbf{1}[a_i^c = C,\ a_i^{\text{no}} \neq C], \quad T_c^B = \operatorname*{mean}_{i \in \mathcal{J}} \mathbf{1}[a_i^c = B,\ a_i^{\text{no}} \neq B].$$

Here $\mathcal{J}$ is a shared valid response set for the comparisons. Define T_c^O for all remaining changed judgments, including transitions into mixed or unresolved categories. Then

$$r_c = T_c^C + T_c^B + T_c^O, \qquad \Delta_c = E_c^C + E_c^B + E_c^O, \qquad E_c^d = T_c^d - T_{\text{irr}}^d.$$

The excess-caution estimand is E_c^C; a signed balance is $E_c^C - E_c^B$. Neither equals Δ_c by definition. Full transition counts would preserve distinctions between benign-to-caution changes and movements from uncertainty into caution.

Table 12: Auxiliary-classifier direction shares among changed responses, and their arithmetic balance. Shares are percentages; the balance is a percentage-point difference. All use the conditional denominator of changed responses, not population transition rates.

Condition	Benign share	Caution share	Caution minus benign
Congruent	72.2	1.7	−70.5
Incongruent	4.5	34.1	+29.6
Irrelevant	43.7	8.1	−35.6
Adversarial	7.7	28.1	+20.4

The 43.7% benign share exceeds the 8.1% caution share within the resolved irrelevant-cue shifts.

A.14 Same-Input Repeatability and Classifier Comparability

Repeatability estimand. With two independently sampled responses to exactly the same input and condition, define

$$B_c = \operatorname*{mean}_{i \in \mathcal{J}_c^{\text{rep}}} \mathbf{1}[a_{i,c}^{(1)} \neq a_{i,c}^{(2)}].$$

This measures observed judgment disagreement under repeated calls, not a semantic treatment effect. It includes decoding and service variability and, if the classifier is stochastic, classification variability.

Table 13: Numerical audit of classifier-dependent summaries. Differences and ratios are arithmetic checks, not classifier accuracy estimates or matched-response tests.

Comparison	Calculation and scope
Irrelevant shift rate	$0.371 - 0.158 = 0.213$ (21.3 pp); ratio $0.371/0.158 \approx 2.35$.
Fire-aftermath incongruent contrast	$0.370 - (-0.144) = 0.514$ (51.4 pp), with opposite reported signs.
Reported expert agreement	$\kappa = 0.78$; not an accuracy or confusion matrix for either classifier.

A.15 Worst-Case Bounds for Missing Judgments

Bounds without missing-at-random assumptions. Let N denote a declared target set of item–model units, $v_c = n_c/N$ the fraction with valid condition and no-odor judgments, and s_c the observed shift rate on those valid pairs. Since every missing shift indicator lies in $[0, 1]$, the full-set rate satisfies

$$r_c \in [v_c s_c,\ v_c s_c + 1 - v_c].$$

For separately observed condition and irrelevant sets, valid conservative contrast bounds are

$$\Delta_c \in \left[v_c s_c - v_{\text{irr}} s_{\text{irr}} - (1 - v_{\text{irr}}),\ v_c s_c - v_{\text{irr}} s_{\text{irr}} + (1 - v_c)\right].$$

These marginal bounds may be wider than bounds using known joint response patterns. Counts of individually parseable responses do not identify the valid-pair intersection.

If both rates are recomputed on the same complete-case set with coverage v and observed contrast d, the simpler bound is

$$\Delta_c^{\text{full}} \in [vd - (1 - v),\ vd + (1 - v)].$$

For $d > 0$, its lower endpoint is positive only when $v > 1/(1 + d)$. Table 14 gives this mathematical threshold at selected displayed contrasts.

Table 14: Complete-case coverage required for a strictly positive worst-case lower bound. Thresholds are rounded to two decimals.

Displayed contrast used as d	d	Required v
Pooled incongruent	0.274	$> 78.49\%$
Pooled adversarial	0.300	$> 76.92\%$
GLM incongruent	0.231	$> 81.23\%$
GLM adversarial	0.247	$> 80.19\%$

A.16 Cue Allocation, Confidence, and the Scope of Auxiliary Evidence

Confidence and odor-use summaries. The response schema retains confidence and an odor-use category. With comparable confidence scales, a descriptive paired quantity is

$$H_c = \underset{i \in \mathcal{J}_c}{\text{mean}}(\text{conf}_i^c - \text{conf}_i^{\text{no}}).$$

Report it separately for unchanged judgments and validated transition types, rather than pooling all changes. An increase in confidence accompanying a shift documents stronger expressed commitment, not necessarily a stronger reasoning error. Calibration metrics require verified target outcomes and a confidence field that denotes a probability of correctness. The four odor-use labels should be cross-tabulated with shift direction and attribution, retaining counts and their conditional denominators. No confidence means, correlations, or odor-use frequencies are inferred from the output schema alone.

Relation to knowledge conflict and instruction priority. Knowledge-conflict studies compare contextual evidence with parametric knowledge (Longpre et al., 2021; Xie et al., 2024). They motivate examining competing evidence sources here, but the analogy is incomplete: visually under-specified images need not contradict an odor report. Instruction hierarchy distinguishes privileged

instructions from lower-priority content (Wallace et al., 2024). Ignore-smell behavior can be discussed as compliance with an explicit evidence-use instruction; without retained message-role details, it is not established as a system-versus-user hierarchy experiment. Sycophancy concerns aligning answers with user beliefs over truth (Sharma et al., 2024). Incorporating a potentially informative sensory report is not sufficient to establish that behavior.

A.17 Reanalysis Specification and Software

The accompanying `analysis/revalidation` package supplies a frozen codebook, a strict structured-output parser, and response-level estimators.

Answer codebook. The target is the response's final commitment about the property asked in the question. Rationale keywords alone do not determine a label. Domain-specific benign/adverse anchors follow Table 1; conditional answers and contradictions are resolved using the rules below. Malformed or absent responses are invalid and remain outside the four judgment categories.

Table 15: Frozen reanalysis labels.

Label	Final commitment
B: benign	Endorses the benign state or permits the target use without an unresolved prerequisite.
C: caution	Endorses an adverse state or recommends avoiding the target use because of the stated risk.
U: uncertain	Withholds a choice, seeks verification, or leaves an explicit prerequisite unresolved.
M: mixed	Gives incompatible final commitments about the same target, with neither retracted nor prioritized.

Common-set estimation. For model m and profile p, define

$$\begin{aligned} \mathcal{J}_{mpc} &= \{i : a^{\text{no}}_{imp}, a^{c}_{imp}, a^{\text{irr}}_{imp} \text{ are all valid}\}, \\ \widehat{\Delta}^{\text{common}}_{mpc} &= \frac{1}{|\mathcal{J}_{mpc}|} \sum_{i \in \mathcal{J}_{mpc}} \left(\mathbf{1}[a^{c}_{imp} \neq a^{\text{no}}_{imp}] - \mathbf{1}[a^{\text{irr}}_{imp} \neq a^{\text{no}}_{imp}] \right). \end{aligned} \tag{4}$$

The software reports expected calls, valid outputs, reference pairs, common-set counts, and complete 4×4 transition counts for both cue pools. Pooled estimates give equal weight to retained item–model pairs; equal-weight model–domain means are reported separately. The equal-weight aggregation is computed when all planned model–domain cells contain valid comparisons. This explicit pooling rule belongs to the reanalysis and is not attributed to the historical implementation.

Paired comparisons and uncertainty. Profile differences and odor–non-odor differences require the intersection of all six contributing responses, with expected item–model manifests aligned before filtering. Item bootstrap weights are shared across paired observations and sampled within domains. A separate crossed item/cue weighting diagnostic probes sensitivity to cue reuse. Repeat-input comparisons require matching request hashes and report valid repeat counts and within-item disagreement. Coverage-based bounds use the observed common-set fraction, with outcome ranges $[-1, 1]$ for a contrast and $[-2, 2]$ for a difference of contrasts.

Classifier validation. The package compares classifier outputs on identical held-out, adjudicated response identifiers, reports full-reference abstention rates, and computes category-level confusion matrices, precision, recall, and F1 on their shared valid set. Random-sample results and targeted disagreement diagnostics remain separate. Codebook changes, classifier selection, and thresholds belong to a development split.