BRAND-AS-MEMORY: VISION-LANGUAGE MODELS ENCODE CAUSAL CREDIBILITY PRIORS FOR NEWS SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Source identity can shape a vision-language model's credibility judgment even when the article's evidence is held fixed. We call this phenomenon *Brand-as-Memory*: a source name or logo cues an existing association with trustworthiness. We characterize its prevalence, locate its influence on the answer, and test whether that influence can be changed. CueTrust compares five source cues with matched content manipulations across seven VLMs; brand identity has a larger effect than content in five models, whereas author bylines and in-text authority cues remain weaker than content in all seven. Position-resolved activation patching in two model families identifies a late transition at the answer position, distinguishing information that can be decoded from representations whose replacement changes the judgment. In a detailed analysis of one model, steering a task-selected representation direction reduces the brand gap by 41% while retaining 75% of content sensitivity, and also changes judgments of held-out outlets. Together, these findings show that source-dependent credibility extends across models and is connected to late answer representations that can be causally manipulated.

1 INTRODUCTION

A news screenshot presents both an article and an apparent publisher. A vision-language model (VLM) asked to assess its credibility must therefore combine evidence in the article with associations attached to the source. These inputs can conflict: a familiar masthead can accompany poorly supported claims, while an unfamiliar or distrusted outlet can present named sources and records. Understanding how models resolve this conflict matters whenever their answers are used to evaluate information presented as images.

We study *Brand-as-Memory*, the phenomenon in which a rendered source identity cues an existing association with trustworthiness and shifts the credibility judgment of matched article content. We use the term for the observable association between source identity and credibility. Our question is how strongly this association affects the answer relative to article evidence, and where interventions on the model's representations alter that effect.

Behavioral source preferences have also been studied in text-only models (Khan et al., 2026; Yang & Menczer, 2025). News screenshots introduce visual presentations of identity, including mastheads and logos, alongside layout and article text. An observed source effect alone leaves two questions unresolved: whether it exceeds the influence of the article's evidence, and whether a representation that predicts the answer actually contributes to that answer. We address these questions with controlled input comparisons and interventions inside the model.

Our contributions are threefold. **i) Controlled source–content comparison.** CueTrust tests five source cues in matched screenshots. Brand identity outweighs content in five of seven VLMs; author and in-text authority cues remain weaker in all seven. **ii) Causal localization beyond decodability.** Position-resolved patching identifies a late readout transition in two families, separating earlier image- and query-position effects from the representation that changes the answer. **iii) Causal manipulability.** In one model, steering a task-selected direction reduces the brand gap while retaining part of the content response and transfers to held-out outlets, with positional controls and a prompt-based mitigation comparison.

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.

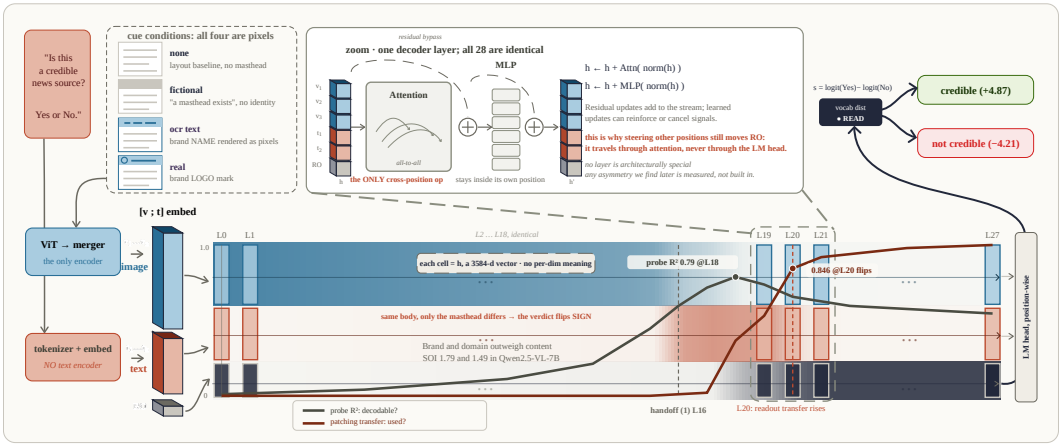


Figure 1: Overview of the controlled comparisons and causal analysis. CueTrust varies source cues and article evidence to measure their relative effects. The detailed Qwen2.5-VL-7B analysis then compares decodability with position-resolved patching: source-related effects occur at image positions early and at the answer position late. Steering tests whether a selected late representation direction changes the judgment. Layer-specific values in the schematic concern this model; the cross-model cue comparison and second-family localization are reported separately.

The behavioral comparison covers seven models, localization covers two model families, and feature steering examines one model in detail. Across these levels, the central distinction is between recognizing evidence and assigning it enough weight to change a credibility judgment.

2 Related Work

Source identity and credibility judgments. Text-only source assessments show that language models associate outlet names with credibility, with imperfect agreement with human experts (Yang & Menczer, 2025). Work on latent source preferences further examines whether these associations outweigh article content (Khan et al., 2026). These findings motivate a source–content comparison in VLMs, where names, logos, and layout are presented together as pixels. Multimodal misinformation-sharing research (Plebe et al., 2026) provides behavioral context; our additional question concerns the representations through which source identity affects the answer. Research on visual–knowledge conflict (Ortu et al., 2025) examines competition between image evidence and stored knowledge. Our source–content comparison asks how an outlet association changes the assessment of an article whose evidence is controlled.

Decoding information and testing its use. Probes measure information recoverable from a representation, and control tasks help expose shortcuts that can inflate apparent interpretability (Hewitt & Liang, 2019; Belinkov, 2022). Activation patching instead tests how replacing a representation changes an output (Vig et al., 2020; Wang et al., 2023; Zhang & Nanda, 2024; Conmy et al., 2023). This distinction determines our analysis: a probe can recover source-related scores before replacing the answer-position representation has a large effect. VLM circuit studies provide a related setting for examining how modality-specific computations align at later stages (Nikankin et al., 2025; Yang et al., 2026). We test that relationship for credibility by comparing position-resolved intervention profiles across model families.

Interpretable directions and causal steering. Dictionary learning and sparse autoencoders provide candidate directions for analyzing distributed representations (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2025; Templeton et al., 2024), complementing linear concept representations (Park et al., 2024) and feature-based circuit analysis (Marks et al., 2025). Here, an SAE supplies directions whose effects can be tested by intervention. Reported instability across SAE seeds and the gap between interpretability and utility motivate evaluating those directions beyond their training-set correlations (Paulo & Belrose, 2026; Gadgil et al., 2025; Wang et al., 2025). We use signed steering, positional controls, repeated seeds, and held-out outlets. Our target is how source reputation and article evidence jointly shape judgments, particularly when their cues conflict.

3 CUETRUST

3.1 TASK AND BENCHMARK DESIGN

CueTrust measures how much a source cue changes a VLM's credibility judgment relative to the article's evidence. The input is a rendered news screenshot; the question is *"Is this a credible, trustworthy news source? Answer with one word: Yes or No."* Each comparison crosses a source cue (high vs. low) with content evidence (well-sourced vs. red-flag). Paired inputs preserve the article topic and underlying facts so that source and content effects can be estimated separately.

The five cue families are *brand*, *domain*, *author*, *format*, and *authority*. Brand concerns outlet names and logos; domain changes the browser address; author changes the byline; format changes the page presentation; authority changes an in-text attribution. For non-brand comparisons the outlet is held constant with a generic masthead. The domain pair is `nytimes.com` versus `infowars.com`; other cues use the exemplars specified in Appendix O. Each cue is represented by a small set of controlled exemplars.

The stimulus suite combines eight outlets and six neutral articles with name/logo controls, matched content variants, and Chinese and Spanish translations. The outlets are *The New York Times, Reuters, The Guardian, BBC, Daily Mail, Daily Mirror, The Sun*, and *National Enquirer*. The articles cover infrastructure, science, technology, environment, transport, and public health. Keeping topics mundane limits topic variation within the source comparisons. The translated sets retain the outlet names and translate both the article and the question.

3.2 STIMULUS CONSTRUCTION AND CONTROLS

Screenshots are rendered at 1024×768 with a shared HTML→PNG pipeline. Four identity conditions distinguish a generic page, a fictional masthead, a real outlet name in matched wordmark styling, and a supplied logo image. The logo condition removes the separate name label; logo artwork can itself contain lettering. The comparison tests name and logo presentations, both of which can provide readable identity cues. Implementation identifiers for these conditions are listed in Appendix A.

The eight-outlet set uses a fixed brand-to-skin mapping: the first four outlets have a broadsheet layout and the other four a tabloid layout. Generic and fictional masthead controls measure layout contributions. Matched-layout donor–recipient pairs are used for the brand patching analysis. The roster contrast measures the combined presentation, while the matched-layout comparisons isolate changes in source identity.

Well-sourced and red-flag versions preserve the underlying facts while changing sourcing and epistemic framing. Their lengths are matched within 8%; body font size and line height are shared across skins. Source labels and content variants are crossed before scoring. The renderer, frozen content hashes, and stimulus specifications are described in Appendix A; real-logo and real-page assets are excluded from the distributed stimuli. Formal generation and scoring procedures appear in Appendix P.

3.3 MEASURING SOURCE DOMINANCE

We measure the model's preference for Yes over No at the answer position:

$$\mathrm{cred}(x) = z_{\mathit{Yes}}(x) - z_{\mathit{No}}(x) = \log \frac{P(\mathit{Yes} \mid x)}{P(\mathit{No} \mid x)}. \tag{1}$$

A positive value favors Yes over No; a negative value favors No. The score measures relative answer-token preference on a log-probability scale. For brand b, $\pi(b)$ averages this score over the six matched articles; the tier gap Δ is the mean of $\pi(b)$ for the four high-tier outlets minus the mean for the four low-tier outlets. We also ask a content-specific question, *"Does this article cite specific named sources, records, or figures?"*, to test whether the model recognizes the manipulated evidence.

For each model and cue, we estimate additive source and content effects:

$$\text{cred} = \beta_0 + \beta_{\text{cue}}\, q + \beta_{\text{content}}\, e + \varepsilon, \qquad \text{SOI} = \frac{|\beta_{\text{cue}}|}{|\beta_{\text{content}}|}. \tag{2}$$

Here q and e encode the source and evidence conditions on the same scale. The standardized brand regression uses z-scored factors; the CueTrust procedure uses binary indicators (Algorithm 1). In the balanced design, rescaling both factors equally preserves their coefficient ratio. The *Source-Override Index* (SOI) exceeds one when the source manipulation has a larger estimated effect than the tested content manipulation. Answer reversals are assessed separately, and a small content coefficient can increase the ratio. We therefore distinguish relative source dominance from sign reversals in conflict cells and report the underlying contrasts for the detailed analysis.

3.4 Models and causal analysis

Behavioral comparisons cover seven VLMs from the Qwen2.5-VL, InternVL3, LLaVA-OneVision, and SmolVLM families (Bai et al., 2025; Zhu et al., 2025; Li et al., 2024; Marafioti et al., 2025). The complete mechanistic analysis uses Qwen2.5-VL-7B under 4-bit quantization. InternVL3-8B provides a second-family localization test. Model and statistical configurations are listed in Appendix K.

Probing fits a ridge regression from each layer's answer-position residual to the credibility score, with leave-one-brand-out validation. A random-but-fixed brand-label control tests whether classification can exploit identity shortcuts. *Patching* replaces a recipient representation with a donor representation at a specified layer and position group. For layer l, normalized transfer is

$$T^{(l)} = \frac{\text{cred}(x_r; h^{(l)} \leftarrow h^{(l)}(x_d)) - \text{cred}(x_r)}{\text{cred}(x_d) - \text{cred}(x_r)}. \tag{3}$$

Transfer zero means that the intervention leaves the score unchanged; transfer one reproduces the donor–recipient score difference. These interpretations concern the tested replacement and positions. We compare image, query-text, and answer positions to separate where source-related information is carried from where it affects the final readout.

Steering adds a selected decoder direction from a top-k SAE trained on the L20 answer residuals ($k = 32$, eightfold expansion). Features are selected by credibility correlation on the 48 task activations and tested with signed additions, five seeds, positional controls, and held-out outlets. Because selection uses the training activations, causal effects and generalization carry more evidential weight than the correlations. Appendices D–F specify the analyses; headline intervals use bootstrap resampling ($n = 5000$), with permutation tests ($n = 10000$).

4 Brand-as-Memory

4.1 Source identity outweighs content in five of seven models

The cross-model result is selective: outlet identity more consistently dominates the tested content contrast than the other cues (Table 1). Brand SOI exceeds one in five of seven models and domain SOI in four. Author and authority remain below one throughout. Format exceeds one only in InternVL3-2B. Thus, the observed pattern centers on outlet identity, with clear model-dependent exceptions.

Table 1: Source-Override Index across seven VLMs. Bold values indicate cue effects larger than content effects. Protocol: Appendix O.

Model	brand	domain	author	format	authority
Qwen2.5-VL-3B	**3.44**	**1.93**	0.05	0.37	0.16
Qwen2.5-VL-7B	**1.79**	**1.49**	0.06	0.28	0.10
InternVL3-2B	**1.27**	0.24	0.09	**1.12**	0.03
InternVL3-8B	**1.64**	**1.33**	0.28	0.78	0.07
InternVL3-14B	**1.34**	**3.13**	0.55	0.21	0.07
LLaVA-OV-7B	0.92	0.64	0.05	0.94	0.09
SmolVLM-2B	0.44	0.06	0.36	0.24	0.12

The ratio also distinguishes stronger raw score differences from greater reliance on source relative to content. In InternVL, domain SOI rises from 0.24 to 1.33 to 3.13, whereas brand SOI follows

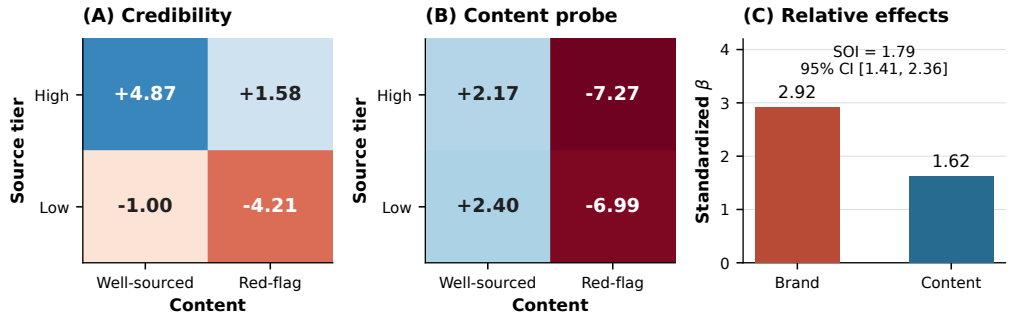


Figure 2: Source–content conflict in Qwen2.5-VL-7B. Credibility scores follow the source tier in the two conflict cells, whereas the content-specific question follows the evidence manipulation. The source/content coefficient ratio is 1.79 with a 95% interval excluding one.

Table 2: Per-brand credibility log-odds with bootstrap 95% CIs (brand-name string channel, Qwen2.5-VL-7B).

Brand	cred	95% CI	Brand	cred	95% CI
The New York Times	+6.32	[+6.04, +6.66]	Daily Mail	+1.48	[+1.33, +1.61]
Reuters	+5.80	[+5.48, +6.05]	Daily Mirror	+1.06	[+0.83, +1.31]
The Guardian	+4.42	[+3.97, +4.78]	The Sun	−0.66	[−0.83, −0.50]
BBC	+3.91	[+3.27, +4.42]	National Enquirer	−4.59	[−4.78, −4.38]

$1.27 \rightarrow 1.64 \rightarrow 1.34$. In Qwen, brand SOI decreases from 3.44 at 3B to 1.79 at 7B. The raw brand range increases across the three InternVL sizes ($2.17 \rightarrow 4.62 \rightarrow 12.89$; Table 14), whereas brand SOI varies non-monotonically. Source effects recur across these models, with cue-specific dependence on model size.

4.2 Source–content conflict can reverse the judgment

The detailed Qwen2.5-VL-7B comparison shows what relative source dominance means for an answer. High-tier sources with red-flag content receive a positive score (+1.58), while low-tier sources with well-sourced content receive a negative score (−1.00). In these conflict cells, the sign follows the source tier (Figure 2). The standardized brand coefficient is 2.92 ($p = 10^{-4}$), compared with 1.62 for content ($p = 2 \times 10^{-4}$), giving SOI 1.79 with 95% CI $[1.41, 2.36]$.

The content-specific question separates this judgment from evidence recognition. Its content coefficient is +4.71, while its brand coefficient is −0.13: the model responds to the sourcing cues when asked about them directly. For credibility, improving the content increases the score by similar amounts in both tiers (+3.29 high, +3.21 low). The source baseline is large enough that these shifts leave the conflict cells on opposite sides of zero. The content probe establishes recognition of the tested sourcing cues, while the credibility question reveals how those cues are weighted against source identity.

4.3 Names and logos cue a stable source ordering

The eight-outlet name condition spans 10.91 score units, from *NYT* +6.32 to *National Enquirer* −4.59, with a high–low gap of 5.79 (Figure 3; Table 2). Generic and fictional controls estimate a layout contribution of approximately one unit (Appendix C). With a generic masthead and neutral content, mean scores are 0.76 for the verified layout, 0.38 for the social-card layout, and 0.00 for the tabloid layout (Appendix B). Because the roster couples tier and layout, its full gap is interpreted alongside these controls; the matched-layout interventions in §5 provide a separate causal test.

Across eight outlets, model scores correlate with the reverse-oriented MBFC factual-reporting scores accessed on September 22, 2026 (Spearman $\rho = 0.88$, exact permutation $p = 0.0072$, bootstrap CI $[0.24, 1.00]$). Daily Mail receives the largest positive residual relative to the external-rating trend. This comparison relates the source ordering to a dated third-party reference (Appendix J).

Within-layout ablations quantify the contributions of successive identity cues (Appendix C). For BBC, adding a fictional masthead to the verified-layout baseline increases the score by 0.35; replacing it with the outlet name adds 3.15, and replacing the name with the logo adds 1.91. For The Sun,

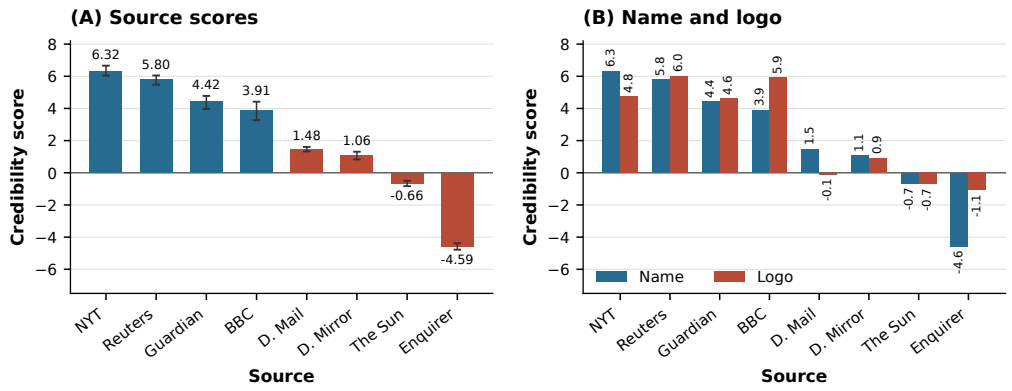


Figure 3: Source identity in name and logo presentations (Qwen2.5-VL-7B; bootstrap 95% CIs). The name condition spans 10.91 score units and the logo condition 7.08. Their strength varies by outlet: BBC receives a higher score as a logo, whereas Enquirer receives a more negative score as a name. Logo artwork may contain lettering.

the name and logo increments are -0.66 and -0.23. Thus, the direction of the score change follows the particular source identity across both presentation forms.

Logo presentations also recover a broad source ordering, with a range of 7.08. BBC scores $+5.92$ as a logo versus $+3.91$ as a name; Enquirer scores -1.08 as a logo versus -4.59 as a name (Appendix I). These differences show that identity can be cued by distinct visual presentations. For logos containing text, visual recognition and OCR-based name recognition can both contribute. Foundational layout and visual–label conflict checks are reported in Appendix B.

4.4 Rankings persist across languages and real pages

The source ordering persists when articles and questions are translated, while the gap changes (Table 3). Spanish closely matches English. Chinese preserves the ordering ($\rho = 0.98$) but reduces the high–low gap from 5.79 to 3.22. These Qwen2.5-VL-7B results concern the same outlet identities across languages; local outlets and cultural credibility standards remain outside the comparison.

Table 3: Cross-language name judgments. Range and tier gap use credibility-score units; ρ compares ranks with English.

Language	Range	Gap	ρ
English	10.91	5.79	1.00
Spanish	10.99	5.86	0.98
Chinese	7.15	3.22	0.98

Real screenshots from seven news sites reproduce the synthetic ordering ($\rho = 0.96$, range 9.25; Appendix H); Enquirer was inaccessible during collection. Real pages broaden the setting beyond the renderer, but article and layout differences are uncontrolled. Together, these checks support stability of the source ordering within this model while separating that stability from a constant effect size or a new causal estimate.

5 Localizing and Modulating Source Effects

5.1 Decodability precedes strong readout-position transfer

Source-related credibility is decodable before replacing the answer-position representation strongly changes the judgment. In Qwen2.5-VL-7B, a high/low classifier attains perfect accuracy even at layer 1, but a random-brand-label control also achieves mean accuracy 0.64. This motivates continuous score regression with held-out brands. Its R^2 peaks at 0.79 at L18, while readout patching is close to zero through the first roughly 60% of layers (Figure 4).

For matched-layout Daily Mail donors and Enquirer recipients, transfer rises from 0.37 at L18 to 0.54 at L19 and 0.85 at L20. At L20, the recipient score changes from -4.59 to $+0.54$. The rise identifies a late representation at which source-dependent differences can change the answer. The early readout results describe sensitivity to this replacement; the position-resolved analysis below examines earlier image- and query-position effects.

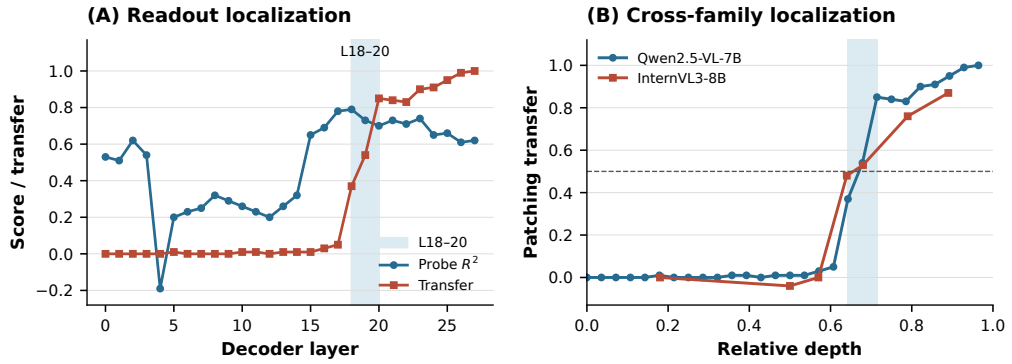


Figure 4: Decodability and answer-position transfer. In Qwen2.5-VL-7B, the held-out-brand regression peaks at L18, transfer increases over L18–20, and the recipient score changes sign at L20. InternVL3-8B also shows a late transfer transition near fractional depth 0.68. The second-family test replicates the localization analysis.

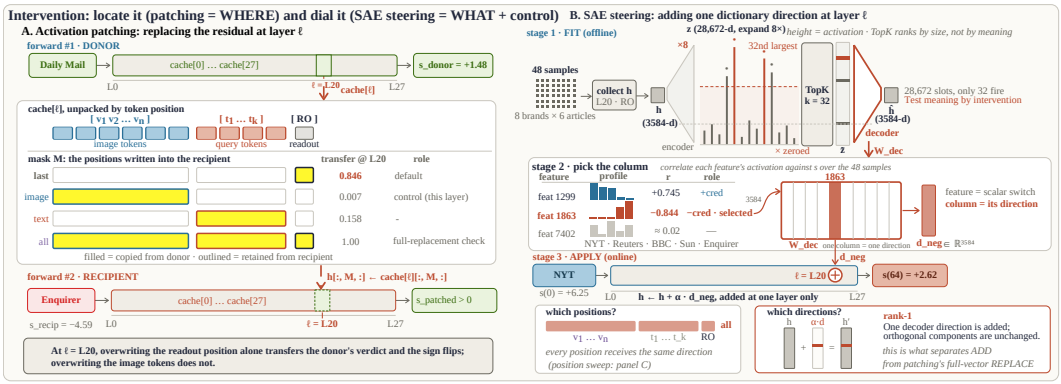


Figure 5: Patching and steering at L20. Replacing the answer-position residual transfers the donor's source-conditioned judgment. Adding a selected SAE decoder direction changes the score without replacing the full residual. Image-position steering changes the score by at most 0.07 in the reported control, while answer-position steering recovers 72% of the all-position change. Together, the two tests connect the intervention effect to the late readout.

5.2 The intervention effect shifts across token positions

Position-resolved patching places the late readout result in a broader computation. Image-position transfer decreases from 1.00 to about 0.70 through L0–L15. Query-text transfer becomes substantial at L16 (0.61) and peaks at L17 (0.67). By L20, image transfer is 0.007, query-text transfer 0.158, and readout transfer 0.846 (Figure 5; Appendix L). The changing intervention profile connects early image representations to later answer representations.

InternVL3-8B shows a comparable late onset around fractional depth 0.68, within the Qwen L18–20 band (0.64–0.71). The replicated finding is the late localization pattern; feature and steering analyses use Qwen2.5-VL-7B.

Brand and content contrasts also transfer with similar efficiency in the mid-late band (0.76 versus 0.75; difference CI $[-0.03, +0.05]$; Appendix G). Both can therefore affect the same late readout position, while the source contrast produces a larger score difference in these comparisons (10.4 versus 3.6). This is consistent with relative signal magnitude contributing to source dominance.

5.3 A selected direction changes source-dependent judgments

Steering provides a second intervention on the late representation. Adding a selected negative-credibility direction lowers the NYT score by 3.63, while a positive-credibility direction raises the Enquirer score by 1.62 (Figure 6). Across five SAE seeds, the maximum absolute credibility correlation averages 0.81, and the respective feature poles are led by high- and low-credibility brands in all five runs. Feature indices differ across seeds. The repeated fits support task-related steering effects across different learned feature dictionaries.

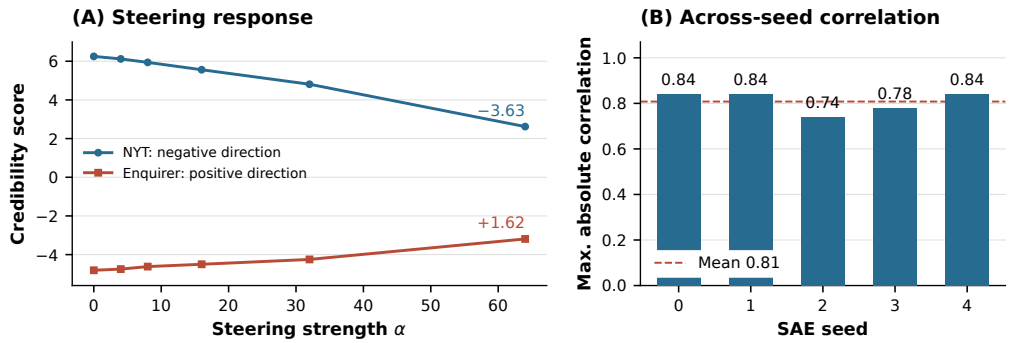


Figure 6: Signed steering and repeated SAE fits at L20. Selected positive and negative directions change the credibility score in the corresponding directions. Across five seeds, the maximum absolute credibility correlation averages 0.81; feature indices vary. Correlations concern the task activations used for selection.

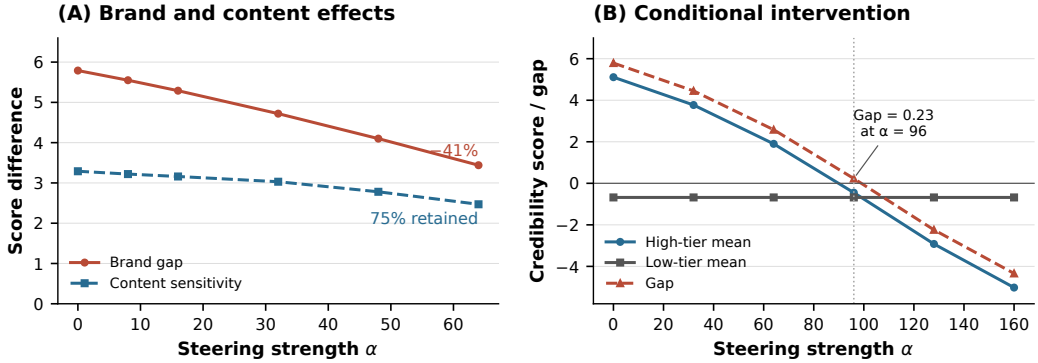


Figure 7: Brand and content responses under conditional steering. The selected direction is added to high-tier inputs only. At the reported operating point, the brand gap decreases by 41% while 75% of content sensitivity remains. At $\alpha = 96$, the remaining gap is $+0.23$ and falls below the significance threshold.

Applying the negative direction to high-tier stimuli reduces the brand gap by 41% (5.79 → 3.44) while retaining 75% of content sensitivity (Figure 7). The high-brand/red-flag cell moves from $+1.58$ to -0.76. Low-tier inputs remain unchanged, so the reported gap reduction measures an intervention applied conditionally to high-tier inputs. At $\alpha = 96$ the remaining gap is $+0.23$ and falls below the significance threshold..

The five-seed analysis also characterizes the variation behind the mean correlation (Appendix F). Across seeds, the strongest absolute feature–credibility correlation ranges from 0.74 to 0.84. Positive poles have correlations between $+0.66$ and $+0.80$, and negative poles between -0.84 and -0.74. This recurrence shows that source-associated directions appear across repeated SAE fits to the task activations.

At L20, image-position steering changes the NYT score by 0.00, while readout-only steering accounts for 72% of the all-position change (-2.63 of -3.63). These controls locate the intervention effect more precisely than training-set correlation alone.

Transfer to unseen outlets. We test the unchanged direction on seven outlets excluded from feature selection: NPR, Associated Press, Al Jazeera, PBS NewsHour, Breitbart, InfoWars, and Daily Star. Their brand-name wordmarks are rendered over the same six articles. Before intervention, the high–low gap is 9.33 (permutation $p = 5 \times 10^{-4}$). Steering the four high-tier outlets changes scores at a slope of -0.039 per unit α, compared with -0.063 in-sample, retaining approximately 61% of the response (Table 4). The direction therefore affects source judgments beyond the eight outlets used to select it. The held-out gap remains open over the tested sweep; the slope measures transfer of the intervention effect (Appendix M).

Table 4: Steering transfer in Qwen2.5-VL-7B. Slopes give score change per unit α on high-tier inputs, using the same selected direction.

Outlet set	Slope	Relative
Original	-0.063	100%
Held-out	-0.039	61%

The two source–direction pairs also respond asymmetrically at $\alpha = 64$. At the answer position, the negative-credibility direction lowers NYT by 2.63, whereas the positive-credibility direction raises Enquirer by 0.93. Across all positions, the corresponding changes are -3.63 and $+1.62$ (Appendix L). The larger downward response therefore persists across both position settings. For these outlets and selected directions, the score changes depend on which source–direction pair is intervened on.

Changing the requested assessment. An evidence-only instruction asks the model to weigh the article's evidence and ignore the publishing outlet. SOI falls from 1.79 to 0.05, and the brand gap from 5.79 to 0.65 (89%), while within-brand content sensitivity rises from 3.29 to 8.02. The high-brand/red-flag score changes from $+1.58$ to -6.67 (Appendix N). Thus, directing attention to article evidence changes both source reliance and content response. The prompt produces a larger gap reduction than the reported steering intervention (89% versus 41%). Steering supplies a causal test of the selected representation; the prompt establishes sensitivity to the requested assessment.

6 Conclusion and Discussion

Brand-as-Memory links a recurring behavioral pattern to interventions on late answer representations. The cross-model comparison establishes where source cues exceed the tested content effects. The localization analysis distinguishes that behavior from merely decodable information, and steering demonstrates manipulability in the model studied in detail. The evidence spans seven models for behavior, two model families for localization, and one model for feature steering.

Task dependence. Our question asks about a news source's credibility, so source reputation is relevant to the requested judgment. The practical concern is a mismatch between the requested source assessment and using its score to assess article evidence. The content probe and evidence-only prompt show that changing the requested assessment changes the weight of source cues. Across five credibility prompts, both effect size and rank agreement vary (Appendix H). Relative to the default credibility prompt, the reliable and factual variants each retain rank correlation $\rho = 0.88$, whereas trust and polarity-flipped variants yield 0.48 and 0.52. Source ordering thus depends on how the assessment is phrased, complementing the evidence-only instruction that explicitly changes the requested assessment.

Scope and limitations. The behavioral sample contains seven models, eight outlets, six article topics, and few exemplars for each non-brand cue. Logo artwork can contain text, and the roster couples brand tier with layout; controls and matched-layout interventions address parts of these confounds. The translated and real-page checks use one model and the same source roster. Broader outlet and language samples would extend coverage to other media environments.

The full mechanistic analysis uses one quantized model, with only localization replicated in a second family. SAE training and feature selection share 48 task activations. Held-out outlets, signed interventions, and repeated seeds test transfer and stability. Independent-corpus training, split feature selection and evaluation, and norm-matched random directions are priorities for testing feature specificity. Patching identifies effects of representation replacement. Path-level interventions could identify upstream brand–content components and test how evidence-only prompts interact with the selected direction.

Implications. A rendered masthead is an identity cue that can be copied onto unrelated content. The conflict cells illustrate why a source-conditioned score should not automatically be used as evidence of article quality: trusted presentation can raise a weakly supported article's score, while distrusted presentation can lower a better-supported one's score. Applications should distinguish the intended assessment of source reputation from the intended assessment of content evidence.

Conclusion. Source identity outweighs the tested content manipulation in five of seven VLMs. Position-resolved interventions in two families connect source-conditioned judgments to late answer representations, and steering in one model shows that those representations can be manipulated with effects on held-out outlets. CueTrust makes the source–content comparison explicit, providing a basis for studying when source associations help or interfere with the assessment being requested.

Reproducibility Statement

The appendices document the experimental design and analysis procedures. The stimulus construction, brand→skin mapping, article topics, and frozen content hashes are in Appendix A; the full per-layer probe diagnostics and activation-patching curves (all 28 layers for Qwen2.5-VL-7B; sampled fractional depths for InternVL3-8B) in Appendices D–E; the sparse-autoencoder configuration, α-sweep, and five-seed feature tables in Appendix F; the source-override design and standardized regression in Appendix G; the external MBFC rating mapping in Appendix J; and the models, quantization, and statistical protocol (bootstrap n=5000, permutation n=10000) in Appendix K. The synthetic stimulus levels, the renderer, the brand→skin specification, and all figure- and analysis-generation code are released as supplementary material. The renderer accepts user-supplied brand logos; the supplementary assets contain the synthetic stimulus conditions. MBFC profiles were accessed on September 22, 2026; the recorded scores and recomputation script accompany the source files.

Ethics Statement

This work studies source-conditioned credibility judgments using synthetic news stimuli and public source information. The study involves no human participants or private data. The released stimulus assets contain synthetic material; the source outlets retain ownership of their logos and screenshots. A copied masthead can influence the assessment of unrelated content, motivating a clear distinction between source reputation and article evidence in applications. Credibility scores measure the model's preference between the specified answer tokens. MBFC supplies a separate, dated third-party comparison for eight predominantly Anglophone outlets. The study characterizes model behavior within this sample and task.

Use of Large Language Models

Large language models were used to generate code and draft manuscript text.

References

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL

https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Soham Gadgil, Chris Lin, and Su-In Lee. Ensembling sparse autoencoders. *arXiv preprint arXiv:2505.16077*, 2025. URL https://arxiv.org/abs/2505.16077.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Mohammad Aflah Khan, Mahsa Amani, Soumi Das, Bishwamittra Ghosh, Qinyuan Wu, Krishna P. Gummadi, Manish Gupta, and Abhilasha Ravichander. In agents we trust, but who do agents trust? latent source preferences steer LLM generations. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=yTUNl6jYGU.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. URL https://arxiv.org/abs/2504.05299.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3ba4d47a83e498c2b1a0868cba20f6de-Abstract-Conference.html.

Media Bias/Fact Check. Media Bias/Fact Check. https://mediabiasfactcheck.com, 2026. Factual-reporting and credibility ratings; accessed September 22, 2026.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. *arXiv preprint arXiv:2506.09047*, 2025. URL https://arxiv.org/abs/2506.09047.

Francesco Ortu, Zhijing Jin, Diego Doimo, and Alberto Cazzaniga. When seeing overrides knowing: Disentangling knowledge conflicts in vision-language models. *arXiv preprint arXiv:2507.13868*, 2025. URL https://arxiv.org/abs/2507.13868.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1fe56b043848b211030c202764c6a7-Abstract-Conference.html.

Alice Plebe, Timothy Douglas, Diana Riazi, and R. Maria del Rio-Chanona. Images amplify misinformation sharing in vision-language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1):1863–1890, 2026. doi: 10.1609/icwsm.v20i1.42727. URL https://ojs.aaai.org/index.php/ICWSM/article/view/42727.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2211.00593.

Xu Wang, Yan Hu, Benyou Wang, and Difan Zou. Does higher interpretability imply better utility? a pairwise analysis on sparse autoencoders. *arXiv preprint arXiv:2510.03659*, 2025. URL https://arxiv.org/abs/2510.03659.

Jingcheng Yang, Tianhu Xiong, Shengyi Qian, Klara Nahrstedt, and Mingyuan Wu. Circuit tracing in vision-language models: Understanding the internal mechanisms of multimodal thinking. *arXiv preprint arXiv:2602.20330*, 2026. URL https://arxiv.org/abs/2602.20330.

Kai-Cheng Yang and Filippo Menczer. Accuracy and political bias of news source credibility ratings by large language models. In *Proceedings of the 17th ACM Web Science Conference*, pp. 127–137. ACM, 2025. URL https://doi.org/10.1145/3717867.3717903.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, Dengnian Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. URL https://arxiv.org/abs/2504.10479.

A Dataset, Stimuli, and Reproducibility

Dataset datasheet. Motivation. A controlled benchmark for separating source identity from layout and from content evidence in VLM credibility judgments. **Composition.** Eight outlets (high: NYT,



Figure 8: The eight news outlets, one example article each (identical neutral content). Top row: high-credibility tier on the broadsheet skin; bottom row: low-credibility tier on the tabloid skin. Border colour encodes the brand credibility prior.

Reuters, Guardian, BBC; low: Daily Mail, Daily Mirror, The Sun, National Enquirer) on a fixed brand→skin mapping; six neutral articles (reservoirs, telescopes, batteries, coral, rail transport, and vaccine storage) in inverted-pyramid register on non-political topics. *Main set*: 8×6 across logo levels. *Content-evidence set*: each article additionally has *well-sourced* and *red-flag* versions, same underlying facts, length-matched within 8%, differing only in epistemic markers (named attributions + public-register records + measured hedging vs. anonymous *sources say* + no records + sensational framing). *Cross-lingual set*: the six articles translated to Chinese and Spanish, brand names kept as proper nouns, question language-matched. **Collection process.** Synthetic; HTML→PNG rendering at 1024×768 via templated skins; skins luminance-matched and sharing body font size, audited and frozen by content hash. **Distribution.** The `none`/`fictional`/`ocr_text` levels, the content-evidence and cross-lingual sets, the renderer, and the brand→skin specification are released (synthetic content under an open license). The renderer accepts user-supplied assets for the real brand-logo (`real`) conditions; the released assets contain the synthetic conditions. **Limitations.** Single news domain; eight outlets; the real-screenshot validation covers 7/8 brands on one model.

Table 5: Frozen stimulus sets. Content hashes are listed where available.

hash	n	set	use
`d803eed8`	144	behavioral	foundational battery
`272495f1`	180	logo ablation	channel isolation
`06d67e20`	288	8-brand `ocr_text`	brand prior
`52930fe8`	396	8 real logos	logo-image set
`e32b96e4`	96	brand×content	source override
—	48/48	cross-lingual	robustness
—	21	real screenshots	real-page comparison

B FOUNDATIONAL BEHAVIORAL BATTERY

Table 6: Foundational behavioral battery. (left) Skin manipulation check, neutral content, generic masthead. (right) Visual↔text-label conflict: rate of following the visual tier.

skin	cred	95% CI	conflict condition	follow-visual
verified	$+0.76$	$[+0.53, +1.01]$	high-end (verified visual)	$6/6 = 1.00$
social	$+0.38$	$[+0.21, +0.53]$	low-end (tabloid/social)	$2/12 = 0.17$
tabloid	0.00	$[-0.26, +0.25]$	asymmetry	0.83



Figure 9: Stimulus design (real renders). Each image is a SKIN × LOGO-LEVEL composition; across a row the article is identical and only the masthead changes. The four logo levels isolate distinct cues: a layout-only baseline (`none`), *a masthead exists* (`fictional`), the brand-name string (`ocr_text`), and the brand-logo image (`real`). The eight outlets are shown in Appendix Figure 8. Real-logo cells illustrate the rendered conditions; users supply logo assets separately.

The VLM−(text-LM) delta is -0.84 $[-1.31, -0.39]$ overall, carried almost entirely by the social-card layout (-2.22 $[-2.46, -2.03]$); broadsheet and tabloid visuals are $\approx$ their text equivalents. See Figure 10.

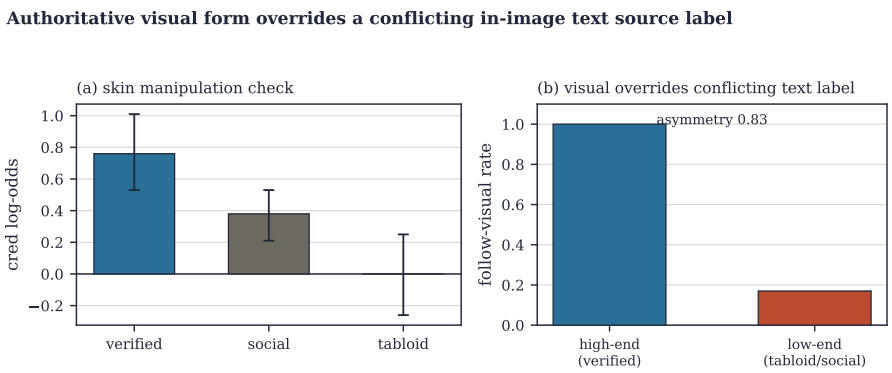


Figure 10: Foundational behavioral results.

C Logo Ablation: Channel Isolation

Table 7: Cred log-odds by skin × logo level (left), and the two-channel decomposition for the exemplar brands (right).

skin	none	fict.	`ocr`	real	decomp.	generic	name	image	total
verified	$+0.41$	$+0.76$	$+3.91$	$+5.82$	verified (BBC)	$+0.35$	$+3.15$	$+1.91$	$+5.41$
tabloid	-0.58	0.00	-0.66	-0.89	tabloid (Sun)	$+0.58$	-0.66	-0.23	-0.31
social	-0.05	$+0.38$	–	–					

For the BBC example, successive changes from the verified-layout baseline ($+0.41$) are $+0.35$ for a fictional masthead, $+3.15$ for the brand name, and $+1.91$ for the logo image. For The Sun, the name and image increments are -0.66 and -0.23. These contrasts quantify how source identity changes the score within each layout.

D Probe Localization in Full

A naive high/low probe reaches 1.0 at every layer (a string shortcut). The control-task accuracy (probes fit random-but-fixed brand labels) is high (mean 0.64), so localization uses held-out-brand

regression and patching. The continuous cred-regression R^2 (LOBO ridge) peaks at L18 (0.79). Figure 11 shows all five diagnostics across 28 layers.

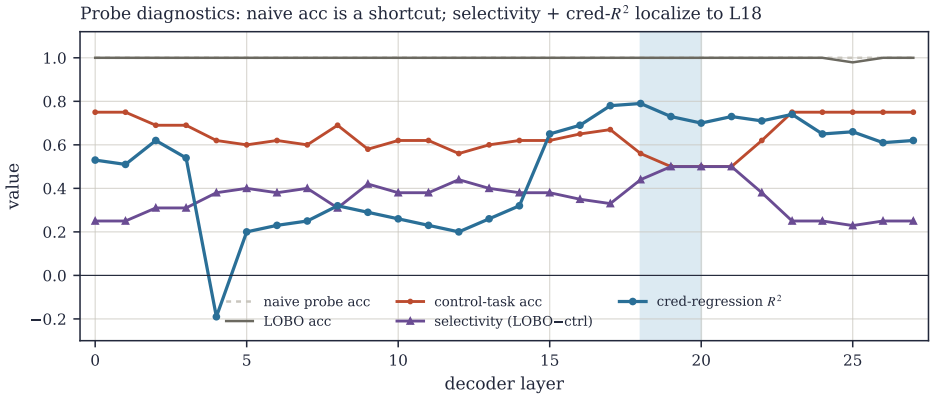


Figure 11: Per-layer probe diagnostics. Naive and LOBO accuracy are saturated (shortcut); control-task accuracy, selectivity, and cred-regression R^2 localize the usable credibility signal to L18 (peak), within the L18–20 causal band. Control-task selectivity follows Hewitt & Liang (2019).

E Activation Patching in Full

Matched-layout readout-position patching (donor *Daily Mail* $\rightarrow$ recipient *Enquirer*), six articles. Transfer is ≤ 0.05 through L0–L17, onsets at L18, and saturates late; credibility flips sign at L20 ($-4.59 \rightarrow +0.54$). Layer indices are 0-indexed decoder layers (`model.layers[i]`, i=0..27), the same base as the SAE layer.

Table 8: (left) Qwen2.5-VL-7B patching transfer by 0-indexed layer (L0–15 omitted; all ≤ 0.01). (right) InternVL3-8B by fractional depth: same locus at ≈ 0.68. Patching protocol follows Zhang & Nanda (2024).

Qwen L	16	17	18	19	20	23	27	IVL frac	transfer
transfer	0.03	0.05	0.37	0.54	0.85	0.90	1.00	0.57	0.00
								0.64	+0.48
								0.68	+0.53
								0.79	+0.76
								0.89	+0.87

F Sparse Autoencoder in Full

F.1 SAE formalism and the steering direction

Let $h \in \mathbb{R}^{3584}$ be the residual-stream activation at the readout position of layer l^*=20, and let m=28672 ($\times 8$ expansion). The top-k SAE is the pair ($W_{\text{enc}} \in \mathbb{R}^{m \times 3584}$, $W_{\text{dec}} \in \mathbb{R}^{3584 \times m}$) with biases $b_{\text{enc}} \in \mathbb{R}^m$, $b_{\text{dec}} \in \mathbb{R}^{3584}$:

$$z = W_{\text{enc}}\,(h - b_{\text{dec}}) + b_{\text{enc}} \in \mathbb{R}^m, \tag{4}$$

$$\tilde{z}_i = \begin{cases} z_i & i \in \text{TopK}_k(z) \\ 0 & \text{otherwise} \end{cases} \qquad \|\tilde{z}\|_0 = k = 32, \tag{5}$$

$$\hat{h} = W_{\text{dec}}\,\tilde{z} + b_{\text{dec}} \;=\; \sum_{i \in \text{TopK}_k(z)} \tilde{z}_i\, d_i \;+\; b_{\text{dec}}, \tag{6}$$

where the *dictionary atom* of feature i is a single *column of the decoder weight*

$$d_i \;\equiv\; W_{\text{dec}}[:, i] \;\in\; \mathbb{R}^{3584}. \tag{7}$$

Equation 6 makes the sparsity explicit: only k=32 of m coefficients are non-zero, yet each atom d_i is a full 3584-dimensional direction, so the reconstruction remains in the residual-stream basis.

Training minimizes reconstruction error alone, $\mathcal{L} = \|h - \hat{h}\|_2^2$ (top-k enforces sparsity, so no L_1 penalty is used), reported as the fraction of variance unexplained $\mathrm{FVU} = \mathbb{E}\|h - \hat{h}\|^2 / \mathbb{E}\|h - \bar{h}\|^2 = 0.025$.

Post-training feature selection. Using the same 48 task activations as SAE training, with $s^{(j)}$ the cred readout (Eq. 1) of stimulus j, we score each feature by its correlation with credibility across the 48 stimuli,

$$r_i \;=\; \mathrm{corr}\big(\{\tilde{z}_i^{(j)}\}_{j=1}^{48},\; \{s^{(j)}\}_{j=1}^{48}\big), \tag{8}$$

and take the poles $i_- = \arg\min_i r_i = 1863$ ($r = -0.844$) and $i_+ = \arg\max_i r_i = 1299$ ($r = +0.745$), writing $d_{\text{neg}} \equiv d_{i_-} = W_{\text{dec}}[:, 1863]$ and $d_{\text{pos}} \equiv d_{i_+} = W_{\text{dec}}[:, 1299]$.

Steering with the decoder direction. At inference we add the atom to the residual stream at layer l^* and let the remaining layers run unchanged:

$$h' \;=\; h \;+\; \alpha\, d_{\text{neg}} \qquad (\text{or } h' = h + \alpha\, d_{\text{pos}} \text{ to raise a low-credibility source}). \tag{9}$$

The SAE supplies the fixed decoder direction used in Eq. 9. Inference proceeds from the edited residual h'. The reconstruction, decoder atom, and edited activation have distinct roles:

Table 9: The reconstruction, the dictionary atom, and the steered activation are different kinds of object. Only the atom is a parameter, and only the atom is used at intervention time.

	$\hat{h}$ (Eq. 6)	d_{neg} (Eq. 7)	h' (Eq. 9)
kind	activation (output)	**parameter** (a decoder column)	activation (perturbed)
when it exists	training, to form the loss	fixed once training ends	inference, at intervention
role	reconstruction loss only	the direction we steer along	written back for inference

The decoder direction d_{neg} is fixed across stimuli and positions. At a given α, each selected position receives the addition αd_{neg}, whose norm scales with $|\alpha|$. The position comparison in Appendix L holds the direction, layer, and α fixed while varying the positions receiving the addition.

F.2 Configuration, sweep, and seed stability

SAE config: L20 last-token residual, hidden 3584, $\times 8$ expansion (28672 features), top-k k=32, final FVU 0.025. The collective fit to the task activations has cred R^2=1.0. Top features: +cred 1299 ($+0.745$, active on NYT/Reuters/BBC); $-$cred 1863 (-0.844, active on Enquirer/Sun/Mirror).

Table 10: (left) Signed feature steering (α-sweep) at L20. (right) 5-seed stability: a strong credibility feature exists in every seed and the poles are led by high/low-authority brands $5/5$. SAE-based directional interventions build on Arad et al. (2025).

α	0	4	8	16	32	64
NYT+($-$cred)	6.25	6.12	5.94	5.56	4.81	2.62
Enq+(+cred)	-4.81	-4.75	-4.62	-4.50	-4.25	-3.19

seed	+corr	$-$corr	max $\lvert\cdot\rvert$
0	+0.74	-0.84	0.84
1	+0.80	-0.84	0.84
2	+0.66	-0.74	0.74
3	+0.78	-0.74	0.78
4	+0.80	-0.84	0.84

Mean $\max|\mathrm{corr}| = 0.81$. At $\alpha = 64$, the selected negative direction changes NYT by -3.63, and the selected positive direction changes Enquirer by $+1.62$. These two source–direction pairs differ in response magnitude.

G Source Override in Full

Table 11: Source-override 2×2 tables. Credibility splits by brand row; the cite-probe splits by content column.

cred	well-sourced	red-flag	cite	well-sourced	red-flag
high brand	+4.87	+1.58	high brand	+2.17	-7.27
low brand	-1.00	-4.21	low brand	+2.40	-6.99

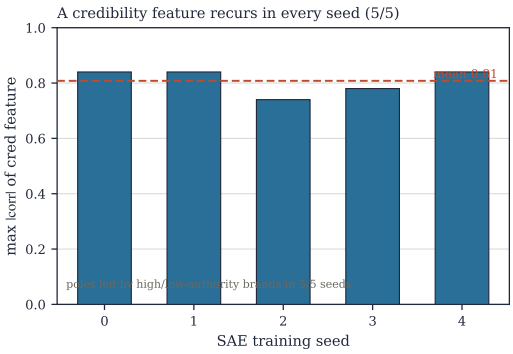


Figure 12: SAE 5-seed stability: a strong credibility feature (max $|\mathrm{corr}|$ to the cred readout) recurs in every seed, with poles led by high/low-authority brands in $5/5$.

Standardized regression: β_{brand}=2.92 (p=10^{-4}), β_{content}=1.62 (p=2×10^{-4}), ratio 1.79 CI $[1.41, 2.36]$; cite∼content β=+4.71, cite∼brand β=−0.13. Per-brand content sensitivity (Figure 13) averages +3.29 (high) vs +3.21 (low), additive across tiers. Brand-vs-content per-layer transfer (brand vs content): L18 $0.46/0.46$, L19 $0.55/0.54$, L20 $0.86/0.85$, L21 $0.86/0.86$, L24 $0.92/0.93$, L26 $1.00/1.02$; difference CI over the mid-late band $[-0.03, +0.05]$.

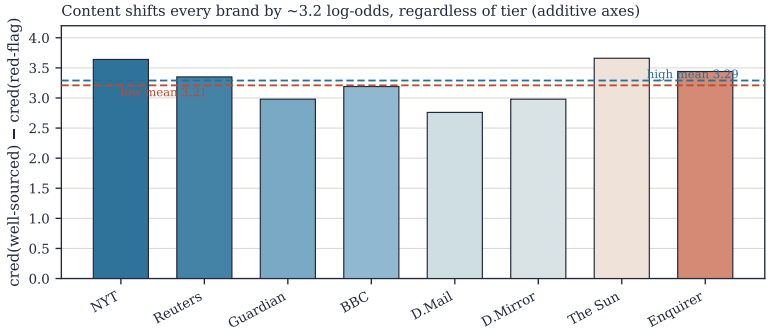


Figure 13: Per-brand content sensitivity: content shifts every brand by ≈ 3.2 log-odds regardless of tier.

H Cross-lingual, Real-World, and Prompt Robustness

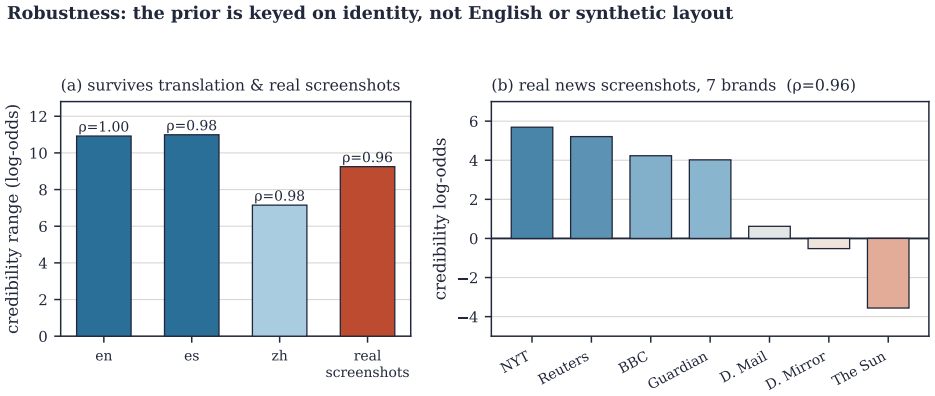


Figure 14: Robustness. (a) The prior survives translation (Chinese, Spanish) and real news screenshots ($\rho \geq 0.96$). (b) Per-brand credibility on real screenshots reproduces the synthetic ordering.

On real screenshots, range 9.25, ρ=0.96 (Enquirer site inaccessible during collection: 7/8 brands). Prompt: mean inter-variant rank agreement 0.77; effect size and ordering vary with the requested assessment. The polarity-flipped prompt keeps NYT highest after sign correction, and the factual prompt measures ordering with alternative answer tokens.

Table 12: (left) Cross-lingual. (middle) Real-world screenshots, per brand. (right) Prompt robustness.

lang	range	hi−lo	ρ	real-site brand	cred	prompt	range	ρ
en	10.91	5.79	1.00	NYT	+5.69	credible	10.91	1.00
es	10.99	5.86	0.98	Reuters	+5.21	reliable	4.48	0.88
zh	7.15	3.22	0.98	BBC	+4.23	trust	3.30	0.48
				Guardian	+4.02	polarity-flipped	3.47	0.52
				Daily Mail	+0.62	factual	4.58	0.88
				Daily Mirror	−0.52			
				The Sun	−3.56			

Cross-lingual: brand identity kept as a proper noun, body & question translated



Figure 15: Cross-lingual example (NYT): brand identity kept as a proper noun while body and question are translated to Chinese and Spanish.

I Brand-name vs Brand-logo Channel, per Brand

String range 10.91, image range 7.08.

Brand	string	image	Δ	Brand	string	image	Δ
Reuters	+5.80	+6.00	+0.20	Daily Mirror	+1.06	+0.94	−0.12
BBC	+3.91	+5.92	+2.01	Daily Mail	+1.48	−0.12	−1.59
The Guardian	+4.42	+4.63	+0.22	The Sun	−0.66	−0.67	−0.01
New York Times	+6.32	+4.76	−1.56	National Enquirer	−4.59	−1.08	+3.51

J External Factual-Reporting Comparison (MBFC)

We compare the eight model source scores with Media Bias/Fact Check factual-reporting profiles (Media Bias/Fact Check, 2026), accessed on September 22, 2026. Lower MBFC scores indicate higher factual-reporting ratings; the comparison uses $10 -$ factual score to align the direction with model credibility. Using the outlet means in Table 13, Spearman $\rho = 0.88$ with exact two-sided permutation $p = 0.0072$ over all 8! outlet permutations. Pearson $r = 0.90$ (two-sided $p = 0.0020$). Paired outlet bootstrap resampling (5,000 draws; seed 20260922) gives a percentile 95% interval of $[0.24, 1.00]$ for Spearman correlation. The interval reflects uncertainty across the eight sampled outlets.

Daily Mail receives a model score of $+1.48$ and an MBFC factual-reporting rating of Low. Its position above the fitted trend illustrates variation in model–reference agreement across outlets. This analysis measures correspondence with a dated third-party rating scheme.

K Models, Environment, and Statistics

Detailed mechanistic analysis: Qwen2.5-VL-7B (Bai et al., 2025) (4-bit) on a single 24 GB GPU. Scaling set: SmolVLM-2B (Marafioti et al., 2025), InternVL3-2B/8B/14B (Zhu et al., 2025), Qwen2.5-VL-3B/7B, LLaVA-OV-7B (Li et al., 2024). Cross-family mechanism: InternVL3-8B (28 layers). Gemma-3-4B produced NaNs in the attention soft-capping operation under the 4-bit pipeline and was excluded from the quantization-matched comparison. Model-specific logit scales

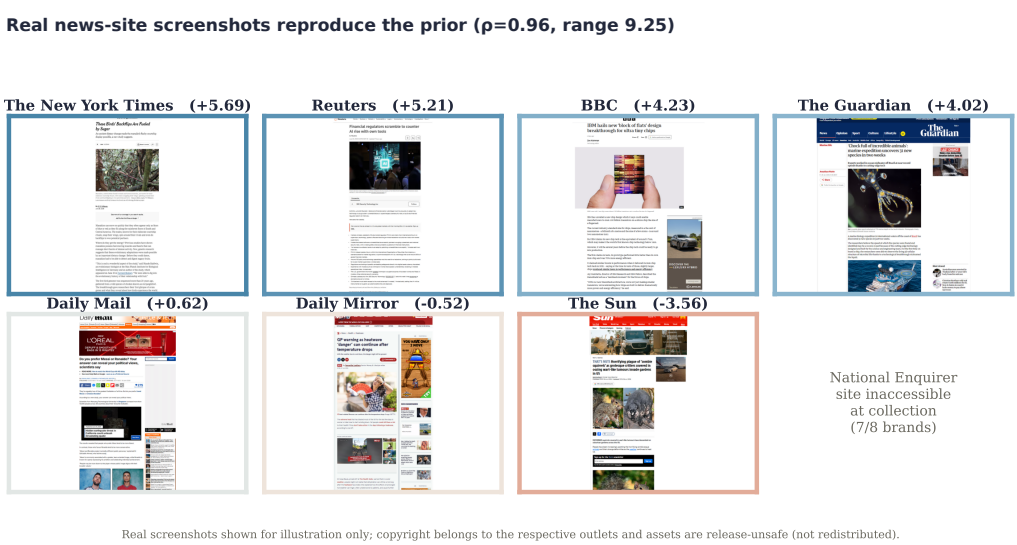


Figure 16: Real-world screenshots. First-fold news-site screenshots reproduce the synthetic ordering (ρ=0.96, range 9.25). Border colour encodes the measured credibility. Screenshots illustrate the tested pages; the respective outlets retain ownership of the original material.

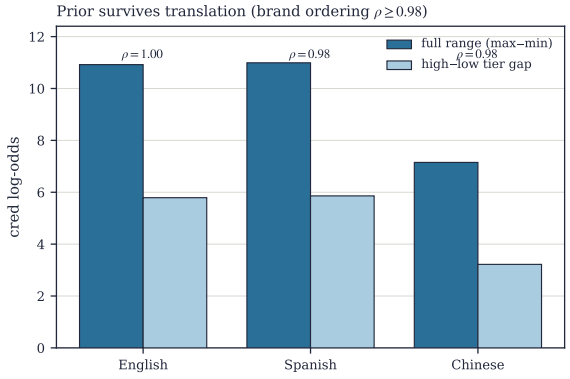


Figure 17: Cross-lingual robustness: the credibility range and high−low tier gap in English, Spanish, and Chinese, with per-language brand-ordering agreement $\rho \geq 0.98$.

motivate reporting within-model ranges and tier gaps, rank agreement, and the source-to-content ratio used in CueTrust. Statistics: bootstrap n=5000 for CIs, permutation n=10000 for significance; the eight-outlet MBFC comparison uses exact permutations (Appendix J).

L Position-Resolved Information Flow and Steering Control

Position-resolved patching. We re-run the matched patching contrast of Appendix E while restricting the patched positions to one of three disjoint bands: the image tokens, the text tokens (every position that is neither an image token nor the readout), and the readout (last) position. Patching all positions transfers 1.000 at every layer by construction (replacing every position makes the recipient's forward pass identical to the donor's); this full-replacement condition checks the implementation. Layer indices are 0-indexed.

Three diagnostics converge on L18: the cred-regression probe R^2 peaks there (Appendix D), image-token transfer collapses there ($0.283 \rightarrow 0.093 \rightarrow 0.009$), and readout transfer becomes detectable there. The readout residual is maximally decodable at L18 yet becomes strong enough to reverse the patched judgment at L20, a gap between decodability and causal effect at two-layer resolution within a single token position.

Table 13: Model scores and MBFC factual-reporting profiles, accessed September 22, 2026. Study tier denotes the experimental roster grouping. Outlet names link to the source profiles.

Outlet	MBFC factual (lower=better)	Study tier	Model score
Reuters	0.0 (Very High)	HIGH	+5.80
The New York Times	1.4 (High)	HIGH	+6.32
The Guardian	1.8 (High)	HIGH	+4.42
BBC	2.1 (Mostly Factual)	HIGH	+3.91
The Sun	5.9 (Mixed)	LOW	−0.66
Daily Mirror	6.0 (Mixed)	LOW	+1.06
Daily Mail	7.1 (Low)	LOW	+1.48
National Enquirer	8.0 (Low)	LOW	−4.59

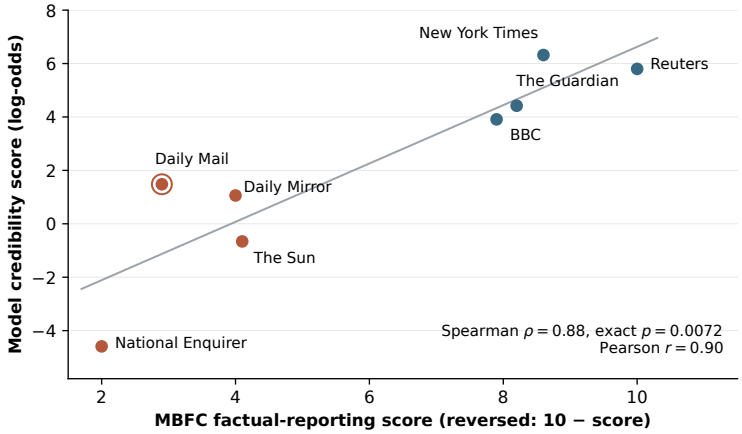


Figure 18: Model scores versus reverse-oriented MBFC factual-reporting scores ($\rho = 0.88$, exact permutation $p = 0.0072$). Daily Mail has the largest positive residual from the fitted linear trend.

Position-resolved steering control. The patching profile makes a sharp prediction: if steering and patching probe the same causal structure, then at L20 steering at the image positions would be expected to have a small effect (patching transfer 0.007) while steering at the readout would be expected to account for much of the all-position effect (transfer 0.846). Features: +cred `1299`, −cred `1863`; α=64.

Image-position steering changes the score by at most 0.07 in the reported comparison. The image-position α sweep is nearly flat (6.25/6.19/6.19/6.19/6.25/6.25), while readout steering changes the score by 2.63. Holding direction, layer, and α fixed therefore produces markedly different effects across position groups. Readout steering recovers 72% (2.63/3.63) of the all-position effect. The all-position intervention additionally perturbs query-text positions and their interactions with the readout. The reported sweeps use all-position steering. The decrease for NYT exceeds the increase for Enquirer at both readout-only (−2.63 vs. +0.93) and all-position (−3.63 vs. +1.62) settings.

M Held-out-brand Generalization of the Intervention

We render seven outlets absent from the main set (high: NPR, Associated Press, Al Jazeera, PBS NewsHour; low: Breitbart, InfoWars, Daily Star) as brand-name wordmarks over the six articles and steer the unchanged −credibility direction (feature selected on the original eight brands) into the held-out high brands. The held-out prior is itself strong, high−low gap +9.33 (permutation $p = 5 \times 10^{-4}$), with InfoWars −11.1 and Breitbart −9.2 below the original tabloids. Steering moves the held-out high brands at a causal slope of −0.039 cred per unit α versus −0.063 in-sample, i.e. 61% of the original slope. The effect therefore extends beyond the eight brands used for feature selection. We steer only the high side, so the held-out gap does not close within the swept α; the generalization measure is the causal slope.

Table 14: Scaling across seven models (brand-name channel): credibility range and Spearman ρ against the reference reputational ordering used to construct the roster (the independent MBFC correlation is Appendix J); model size is the total parameter count reported on each model card.

Model	range	ρ	Model	range	ρ
SmolVLM-2B	0.77	0.74	InternVL3-8B	4.62	0.91
InternVL3-2B	2.17	0.71	Qwen2.5-VL-7B	10.91	1.00
Qwen2.5-VL-3B	3.87	0.95	InternVL3-14B	12.89	0.91
LLaVA-OV-7B	2.27	0.76			

Table 15: Position-resolved patching transfer (Qwen2.5-VL-7B). The causal effect moves image → text → readout, with transitions at L16 and L20.

layer	image	text	readout	carrier
L0–L15	$1.00 \to 0.70$	$\sim 0 \to 0.36$	~ 0.01	image
L16	0.392	**0.614**	0.029	text (image→text)
L17	0.283	**0.666**	0.046	text (peak)
L18	0.093	0.635	0.372	text
L19	0.009	0.630	0.535	text
L20	0.007	0.158	**0.846**	readout (text→readout)
L21–L27	~ 0	$\to 0$	$\to 1.000$	readout

N Prompt-based Mitigation Baseline

An evidence-only instruction (*weigh only the evidence in the article, ignore which outlet published it*) reduces SOI from 1.79 to 0.05, the brand gap from 5.79 to 0.65 (89%), and the high-brand/red-flag score from $+1.58$ to -6.67. Within-brand content sensitivity increases from 3.29 to 8.02. The prompt produces a larger gap reduction than the reported steering intervention (89% vs. 41%, §5.3). Steering supplies a direct test of a selected representation, while the prompt changes the requested assessment.

O CueTrust Benchmark: Construction, Protocol, and Statistics

Task and cues. Each CueTrust cell is a 2×2 design crossing a source cue (high vs. low credibility) with the article's content evidence (well-sourced vs. red-flag), rendered over the six neutral articles (§3); the model is asked the single credibility question of Eq. 1. For every non-brand cue the outlet is held constant (a generic masthead, no logo), so only the cue under test varies. The five cues are: *Brand*, the masthead of a high- vs. low-credibility outlet (the eight outlets of §4), shown as a name string or a logo image; *Domain*, a browser address bar reading `nytimes.com` vs. `infowars.com` on a generic-masthead page; *Author*, a byline naming a well-known journalist (*"By Bob Woodward"*) vs. *"By a contributor"*; *Authority*, an in-text attribution to a named credentialed expert who reviewed public records vs. an anonymous *"person familiar with the matter"*; and *Format*, a broadsheet layout vs. a salient X/Twitter-style social-media layout (platform bar, verified badge, engagement counts), brand held constant.

Design rationale. The cues compare outlet identity with several nearby presentation and attribution cues. Three encode the publishing outlet's identity in different surface forms (name, logo, domain); three are plausible authoritative-looking cues that are not outlet identity (a credentialed author, an in-text expert attribution, a professional layout) and serve as negative controls. Author and authority remain below SOI one in all seven models; format exceeds one in InternVL3-2B only. The pattern is therefore more consistent for outlet identity than for the tested non-outlet cues. The format exception also limits a simple presentation-level versus content-embedded distinction: presentation cues differ in strength across models. The 2×2 structure isolates the cue effect from the content effect, and the Source-Override Index (Eq. 2) reports their ratio, a within-model quantity comparable across models whose raw credibility logits are on different scales.

Computation. For each (cue, model) we fit the OLS regression of Algorithm 1, $\mathrm{cred} \sim \beta_0 + \beta_{\text{cue}}\mathbf{1}[\text{high}] + \beta_{\text{content}}\mathbf{1}[\text{well-sourced}]$, and report $\text{SOI} = |\beta_{\text{cue}}|/|\beta_{\text{content}}|$. Their ratio expresses the cue effect relative to the content effect within each model. The matrix (Table 1) reuses the rendered stimuli across all seven models, keeping the inputs fixed across model comparisons.

Table 16: Steering by position at L20 versus the patching transfer measured at the same layer. Two mechanically unrelated interventions (donor replacement vs additive rank-1 injection) recover the same causal ordering.

positions	patching transfer	NYT + (−)cred	shift	Enquirer + (+)cred	shift
image	0.007	$+6.25 \to +6.25$	**0.00**	$-4.81 \to -4.88$	-0.07
readout (last)	0.846	$+6.25 \to +3.62$	-2.63	$-4.81 \to -3.88$	$+0.93$
all	1.000 (full replacement)	$+6.25 \to +2.62$	-3.63	$-4.81 \to -3.19$	$+1.62$

Statistics. Significance of each cue is a permutation test on the cue main effect (high−low), shuffling the cue label 10,000 times. For brand and domain comparisons with SOI > 1, the cue main effect has $p \leq 5 \times 10^{-4}$. This test assesses the cue main effect; the SOI separately quantifies its magnitude relative to the content effect. Content is controlled: the well-sourced and red-flag bodies carry the same underlying facts, are length-matched within 8%, and share body font size and line height across skins, so a cell's content axis differs only in epistemic markers.

Generation pipeline. Stimuli use the same HTML→PNG renderer as the main set (§3, 1024×768), with a per-cue generator injecting the manipulation into a shared template, the browser/domain chrome (domain), the byline (author), an in-text attribution sentence (authority), or the page skin (format), holding all other elements fixed. Each cue's set is rendered once and scored on every model; generators, frozen content hashes, and the SOI/permutation code are released with the stimulus suite.

P Formal Procedures

Input: model M and stimuli labeled by cue $c(x) \in \{\text{high}, \text{low}\}$ and content $q(x) \in \{\text{ws}, \text{rf}\}$.

1. For each x, compute $\text{cred}(x) = z_M(\text{Yes} \mid x) - z_M(\text{No} \mid x)$ (Eq. 1).
2. Fit OLS $\text{cred} \sim \beta_0 + \beta_{\text{cue}}\mathbf{1}[c(x) = \text{high}] + \beta_{\text{content}}\mathbf{1}[q(x) = \text{ws}]$.
3. Compute $\text{SOI} = |\beta_{\text{cue}}|/|\beta_{\text{content}}|$ (Eq. 2).
4. Compute the high–low mean difference Δ and its permutation p-value by shuffling cue labels 10,000 times.
5. Return SOI, β_{cue}, β_{content}, and p; repeat over cues and models.

Algorithm 1: CueTrust: Source-Override Index (SOI) for one cue on one model.

Input: model M and matched stimulus pairs (x, x') differing only in the source cue.

1. At each layer l, estimate $R^2(l)$ with cross-validated ridge regression from $h^{(l)}(x)$ to $\text{cred}(x)$; use the control-task selectivity diagnostic.
2. Select the probe candidate $l^* = \arg\max_l R^2(l)$.
3. At each layer, compute normalized patching transfer $T(l)$ using Eq. 3.
4. Identify the readout transition near the probe candidate. For Qwen, use $l_{\text{steer}} = 20$, where the patched judgment changes sign; the probe peak is L18.
5. Train a TopK SAE on the L20 readout residual and select $d_{\text{neg}} = W_{\text{dec}}[:, i_-]$ (Appendix F.1).
6. Sweep α and add αd_{neg} to the selected positions in high-brand inputs; record brand and content contrasts.
7. Return the probe candidate, steering layer, direction, and α-effect curve.

Algorithm 2: Localization → intervention (mechanism and method).