# ELM, Disrupted: Large Language Models That Recognize Advertising Evidence Without Using It


## Abstract

The marketing funnel is eroding: large language models (LLMs) now sit between brands and consumers at every stage — selecting creative in programmatic systems, pretesting agency work, and recommending products in chat. But fluency is not judgment, and what moves these decisions goes unexamined. Study 1 synthesizes decades of human advertising evidence (3,276 included reports). Stronger arguments outperform weaker ones on average (eight-contrast reference pool, $r = .41$, 95% CI [.28, .53]), but heterogeneity is the finding ($I^2 = 82\%$; prediction interval [.02, .70]; confirmed across 44 pooled analyses). Study 2 stress-tests five fixed open-weights models (a 22,800-trial controlled grid over novel ad families plus 12,320 natural-ad recognition trials), separating recognition from reliance. The models recognize ad content at 91.6% pooled accuracy (five completed batteries), yet stronger evidence robustly prevails in 0 of 12 panel families — one model minority passes 4 — while control edits sway them as readily as genuine upgrades, and their stated reasons follow from cited premises in 8– 34% of derivable trials. Wherever AI now evaluates ads, fluency without diagnostic sensitivity is the appearance of judgment, not judgment.



*Keywords:* advertising effectiveness; elaboration likelihood model; large language models; computational behavioral testing; meta-analysis; marketing funnel

# Introduction

The marketing funnel assumed a human at every stage: a consumer who notices, considers, and chooses. That assumption is eroding. Large language models now see and evaluate advertisements wherever judgment used to be human — inside programmatic systems assembling and selecting dynamic creative (Bakpayev et al., 2022), inside agencies using AI to create and pretest work (Van Berlo et al., 2024), and inside consumer-facing chat reading marketing content and suggesting winners (Huang & Rust, 2024; citation removed for blind review). At each post, the model answers with unsettling fluency — claims identified, creative summarized, alternatives ranked, a brand recommended — and that fluency passes for judgment (Campbell et al., 2022): budgets shifted, creative killed, products recommended, the empirical question unasked. What such a system responds to — which advertising information moves its task-contingent output, under which conditions, how reliably — remains almost entirely unmapped. Every automated decision skips that question, and the skipping is becoming infrastructure.

The elaboration likelihood model (ELM) offers the discipline this moment needs. ELM links conditions of processing, properties of persuasive communication, and evaluative consequences (Petty & Cacioppo, 1986; Petty et al., 1983). Its central warning: no ad feature is intrinsically central or peripheral — a source cue, image, appeal, or claim serves different functions by recipient, task, and context, and no rating, thought listing, or confidence judgment alone licenses an inference to elaboration. Four decades of advertising research vindicate that conditionality: effects vary, sometimes wildly, across features, conditions, and outcomes. As AI takes funnel posts, that conditionality becomes the audit standard: any system standing in for a human reader must show diagnostic information moves it and cosmetic changes do not — not have fluent output relabeled as a mental state.

Study 1 assembles the human advertising literature into feature $\times$ condition $\times$ outcome evidence distributions — what was varied or measured, for whom and under what conditions,

with what result and what heterogeneity. Study 2 interrogates named, versioned LLMs with thousands of controlled advertising tests: not whether models "are like consumers," but which information moves a model's task-contingent output, under which conditions, how reliably. Functional ELM generalizes the discipline itself, restating ELM's conditional structure as a substrate-independent theory of evaluation, testable by manipulation rather than introspection.

## Theory and Hypotheses

ELM explains persuasion in terms of the extent and nature of issue-relevant thinking. Motivation, ability, and opportunity can alter whether recipients scrutinize information; cognitive responses then shape evaluations and their strength (Petty & Cacioppo, 1986). When conditions favor elaboration, the merits of proposition-diagnostic information should matter more. Otherwise, source, consensus, affect, and executional features can influence judgment through other routes or functions. The function of a feature must therefore be established in a condition.

The relevant ELM chain has five linked elements: what governs scrutiny, what gets scrutinized, how processing shows, what follows, and what qualifies every link.

| Element | Role in the chain | Examples |
|---|---|---|
| Determinants | Govern whether recipients scrutinize | Relevance, need for cognition, knowledge, distraction, time, demand |
| Persuasive inputs | What recipients scrutinize | Claims, argument quality, evidence, source traits, social signals, affect, imagery, narrative, disclosures |
| Diagnostics | How processing shows | Attention, thought direction, scrutiny, counterarguing, credibility, recall, recognition, confidence |
| Outcomes | What follows | Attitude toward the ad, brand attitude, purchase intention, choice, behavior, willingness to pay |

| Element | Role in the chain | Examples |
|---|---|---|
| Boundary conditions | Qualify every link | Product, modality, familiarity, culture, complexity, repetition, timing |

Attitude toward the ad matters but is not interchangeable with brand attitude, product attitude, choice, or behavior (Brown & Stayman, 1992; MacKenzie et al., 1986). Advertising works through intermediate responses — beliefs, attitudes, intentions — whose links to behavior are themselves conditional, so this program keeps outcomes disaggregated. The links shift with message structure: humor lifts brand attitude in one-sided ads but not two-sided ones, where salient negative information captures the attention humor would otherwise deflect (Eisend, 2022).

The argument-quality × involvement interaction is ELM's most replicated prediction. Strong arguments outperform weak ones, and the advantage grows where recipients scrutinize: across 134 studies, the strong-beats-weak advantage runs substantially larger under central than peripheral processing (Carpenter, 2015). In advertising, randomized involvement moves the argument-quality effect on product attitude directly (Petty et al., 1983), and argument-quality contrasts stay positive across involvement conditions (Andrews & Shimp, 1990). Any system standing in for a human reader must show the same conditional signature: proposition-diagnostic information moving output more than matched nondiagnostic cues, stably across tasks — or reveal, by failing to, that its fluency runs on something else.

*H1: Argument dominance versus the task-alternative account (Proposition 1).* Across new paired ad families, strengthening proposition-diagnostic claims or evidence produces larger, more stable shifts in the model's structured evaluative output than matched changes in nondiagnostic source prestige or executional cues. The rival task-alternative account predicts apparent argument dominance shifts, collapses, or reverses when only the decision rule, task wording, output format, or otherwise semantics-preserving frame changes.

*RQ1: Other cue families.* How do social proof, scarcity, celebrity, aesthetic/ executional, affective, narrative, identity, and visual-information cues affect recognition and functional reliance across creative and product families?

The persuasion knowledge model holds that consumers develop lay theories of marketer tactics, deploying skepticism selectively — discounting a celebrity endorsement while crediting verifiable evidence, for example — depending on knowledge accessibility and capacity to use it (Campbell & Kirmani, 2000; Friestad & Wright, 1994). As funnel stages compress around LLM exposure (citation removed for blind review), the question is whether systems at those stages show anything functionally resembling selective resistance. Study 2's credibility tests are built in that image: a system unable to distinguish an evidential link from a prestige label has nothing functionally resembling persuasion knowledge, however fluently it talks about sources.

*H3: Diagnostic credibility versus nondiagnostic prestige (Proposition 3).* Evidence-linked credibility should produce a more robust, evidence-sensitive output change than nondiagnostic prestige. Equal or larger prestige effects under those controls reveal cue-sensitive output — either pattern a functional finding. Hypotheses keep their preregistered labels (H1, H3), matching the preregistration, trial data, and analysis code.

Language models reproduce human response distributions well enough to tempt substitution: silicon samples track survey marginals (Argyle et al., 2023), and marketing guidelines now scope their legitimate uses to collaboration under human oversight (Sarstedt et al., 2024). The critiques bite at the same joint this battery tests: synthetic respondents match opinion averages while compressing within-group heterogeneity and drifting over time (Bisbee et al., 2024), and LLMs fail to reproduce human response biases while swinging with innocuous prompt and order perturbations (Tjuatja et al., 2024).

Stated reasons compound the problem. Natural-language explanations can be unfaithful to the computation producing the answer (Jacovi & Goldberg, 2020; Turpin et al., 2023), so a

fluent rationale is evidence of nothing until an ablation backs it. Recognition and reliance therefore need separate tests — and models already pass the first kind: zero-shot LLM annotation beats crowd workers on text classification at higher agreement (Gilardi et al., 2023).

*RQ2: Recognition, reliance, and rationale (Proposition 4).* When a model can identify a claim or cue, does causally removing or altering that feature change its decision? When it says the feature mattered, does the ablation support that explanation?

ELM's conditional structure survives the removal of the human: its core distinctions restate as properties of *any* evaluation system — person, panel, or program — testable by manipulation rather than introspection. Call the restatement Functional ELM: ELM's discipline without its psychology, one audit logic for every funnel post AI occupies. (1) Evaluators differ in diagnostic robustness, and the difference is measurable. Elaboration becomes a behavioral gap: how much more proposition-diagnostic manipulations move a system's output than matched nondiagnostic ones, stably across tasks. (2) Feature-function is established per system by manipulation. Central and peripheral stop naming mental routes and start naming what ad content *does* in a specified evaluator — moving output under what conditions. (3) Selective resistance is the observable form of persuasion knowledge. (4) Representational elaboration without functional elaboration is detectable — and diagnostic. Functional ELM attributes no mental states, claims no mechanism, and travels only with its tests — a robustness score means nothing outside the feature families, tasks, and decision rules producing it. H1 tests proposition 1, H3 tests proposition 3, and RQ2 probes proposition 4; proposition 2 is established per execution by the ablation and control results.

# Study 1: Feature × Condition × Outcome Human Meta-Analysis

## Corpus and coding

Study 1 asks what the accumulated human record shows — and how much it varies. It is a broad, source-verified synthesis of advertising effects organized by persuasive feature, recipient/ task condition, processing diagnostic, and evaluative outcome.

Search and citation follow-up yielded 7,331 report records. After deduplication and reconciliation, 5,542 unique records remained; 4,744 full texts were retrievable and readable for screening. The descriptive inventory comprises 3,276 reports meeting inclusion criteria, spanning 5,680 studies, 5,661 samples, and 56,603 coded construct instances. The fully source-verified quantitative file holds 16 contrasts from 10 reports, including the eight-contrast reference synthesis.

The content-analysis schema records determinants, persuasive inputs, processing diagnostics, outcomes, boundary conditions, and proposed LLM observability. Image-direct observation occurs only 188 times — an asymmetry that sets Study 2's access design across claims, cues, and images. AI-assisted extraction supported screening triage and numerical candidate generation; every retained value was checked against its source page and statistic before pooling.

## Reference synthesis

Pooling runs at the feature × condition × outcome subgroup level. Where supported, random-effects models, sensitivity analyses, and prediction intervals express the distribution of human results (Eisend, 2017). Subgroups lacking compatible metrics or source evidence remain evidence maps — structured summaries of what was studied — or narrative syntheses.

Only qualifying subgroups were pooled.

Eligible contrasts had to: (a) expose human recipients to an advertising stimulus; (b) compare stronger versus weaker proposition-diagnostic information or informational versus exemplar evidence; (c) report a one-degree-of-freedom $F$ or $t$ statistic for an evaluative outcome; (d) specify a defensible direction; and (e) occupy a unique sample/ dependency group. Signed partial $r$ was calculated from the reported statistic, transformed to Fisher's $z$, assigned sampling variance $1/(df_{error} - 1)$, and synthesized with inverse-variance random-effects REML models. The reference set contains eight independent contrasts from seven reports. Stronger arguments or informational evidence were associated with more favorable evaluative outcomes: random-effects partial $r = .413$, conventional 95% CI [.276, .533], modified Hartung–Knapp sensitivity CI [.246, .556], $\tau_z^2 = .039$, and $I^2 = 82.5\%$. The normal-approximation 95% prediction interval is [.021, .695]. Leave-one-out pooled $r$s remain positive (.375– .450), but heterogeneity is substantial. One additional contrast is held out of all calibration pools pending report-version adjudication and is never pooled.

Two direct within-paper comparisons anchor the conditional logic. In Petty et al. (1983), the argument-quality effect on product attitude was greater under high than low randomized involvement, $F(1, 148) = 71.36$ versus $F(1, 148) = 5.40$, with an involvement × argument-quality interaction of $F(1, 148) = 18.47$. In Andrews and Shimp (1990), argument-quality contrasts were positive under both involvement conditions, although the analysis followed a reported purification step (187 recruited; 117 retained).

## Broad synthesis and heterogeneity

Full extraction covered 2,256 reports, yielding 29,621 extracted effects and 19,819 nonextractable results with documented dispositions. No single number survives the pooling: effects change size — and sometimes direction — with the feature, the conditions, and the outcome. Subgroups followed the preregistered rule: feature × effect × design × outcome,

correlations only, residual "other" categories excluded, minimum 3 effects, 2 dependency clusters, and 2 reports. Forty-four subgroups pooled; 470 subgroups remain evidence maps. Each pooled analysis reports pooled $r$, 95% CI, $\tau^2$, $I^2$ (Higgins et al., 2003), and a 95% prediction interval, with leave-one-cluster-out, Egger regression where $k \geq 10$ clusters (Egger et al., 1997), trim-fill where $k \geq 5$ (Duval & Tweedie, 2000), a within-cluster-correlation sensitivity analysis, and RVE-CR2 with Satterthwaite degrees of freedom where usable (Hedges et al., 2010; Tipton, 2015).

The benchmark tier ($\geq 3$ reports, $\geq 5$ clusters, no leave-one-cluster-out sign flip) admits 10 effects. Peripheral/ heuristic cues predict attitude toward the ad ($r = .403$, 95% CI [.14, .61]). Central/ proposition-diagnostic inputs predict purchase intention ($r = .164$, 95% CI [.02, .30]), but the RVE interval spans zero ([−.03, .27]) — a dependence-sensitive result. Peripheral cues predict purchase intention ($r = .165$, 95% CI [−.03, .35]) with trim-fill attenuation to .045 — a publication-bias-sensitive result. Processing-diagnostic measurement-structure subgroups range $r = .35$– .46. Heterogeneity is high throughout ($I^2$ 89– 98%) and prediction intervals wide. Dependence assumptions move no benchmark finding across zero: pooled means shift at most .008 across the within-cluster-correlation sensitivity analysis in benchmark effects (at most .019 across all 44 pooled analyses).

| Subgroup (effect / construct / design → outcome) | $k$ effects/ clusters/ reports | Pooled $r$ [95% CI] | $I^2$ |
|---|---|---|---|
| Partial / peripheral cue / randomized → attitude toward the ad | 19/ 7/ 6 | .403 [.14, .61] | 98.0 |
| Partial / central input / randomized → purchase intention | 25/ 15/ 7 | .164 [.02, .30] | 95.9 |
| Partial / peripheral cue / randomized → purchase intention | 38/ 24/ 14 | .165 [−.03, .35] | 98.2 |
| Partial / argument quality / randomized → brand attitude | 13/ 6/ 3 | .057 [−.19, .30] | 91.8 |

The remaining benchmark effects cover ability and motivation subgroups and processing-diagnostic measurement-structure subgroups ranging $r = .35$– $.46$. The reference and benchmark-tier forests are plotted in Figure 1.

## From human evidence to machine experiments

Study 1 determines what to test — the features, conditions, and outcomes the human record shows to matter independently — and Study 2 tests each with a precision only re-executable systems allow. Each Study 2 feature family accordingly inherits one of three evidence statuses from the human record: (1) a comparable, source-verified human effect with direction, range, and caveats; (2) a source-verified human evidence map with no commensurable pooled effect; or (3) a new computational test with no asserted human reference. The human record and the machine trials meet at the level of direction. Where the human literature estimates argument-quality associations — positive and small-to-moderate, as in the randomized argument-quality-to-brand-attitude effect ($r = .057$) and the source-verified reference pool ($r = .413$) — the models' H1 forced-choice findings are near-uniformly null.

Study 1's inventory holds few image-direct observations (188 of 9,251 inputs), so which effects need the image stays open. *RQ3: Multimodal and access dependence.* Which effects require text, image, or their joint presentation, and which effects are artifacts of prompt/ task construction?

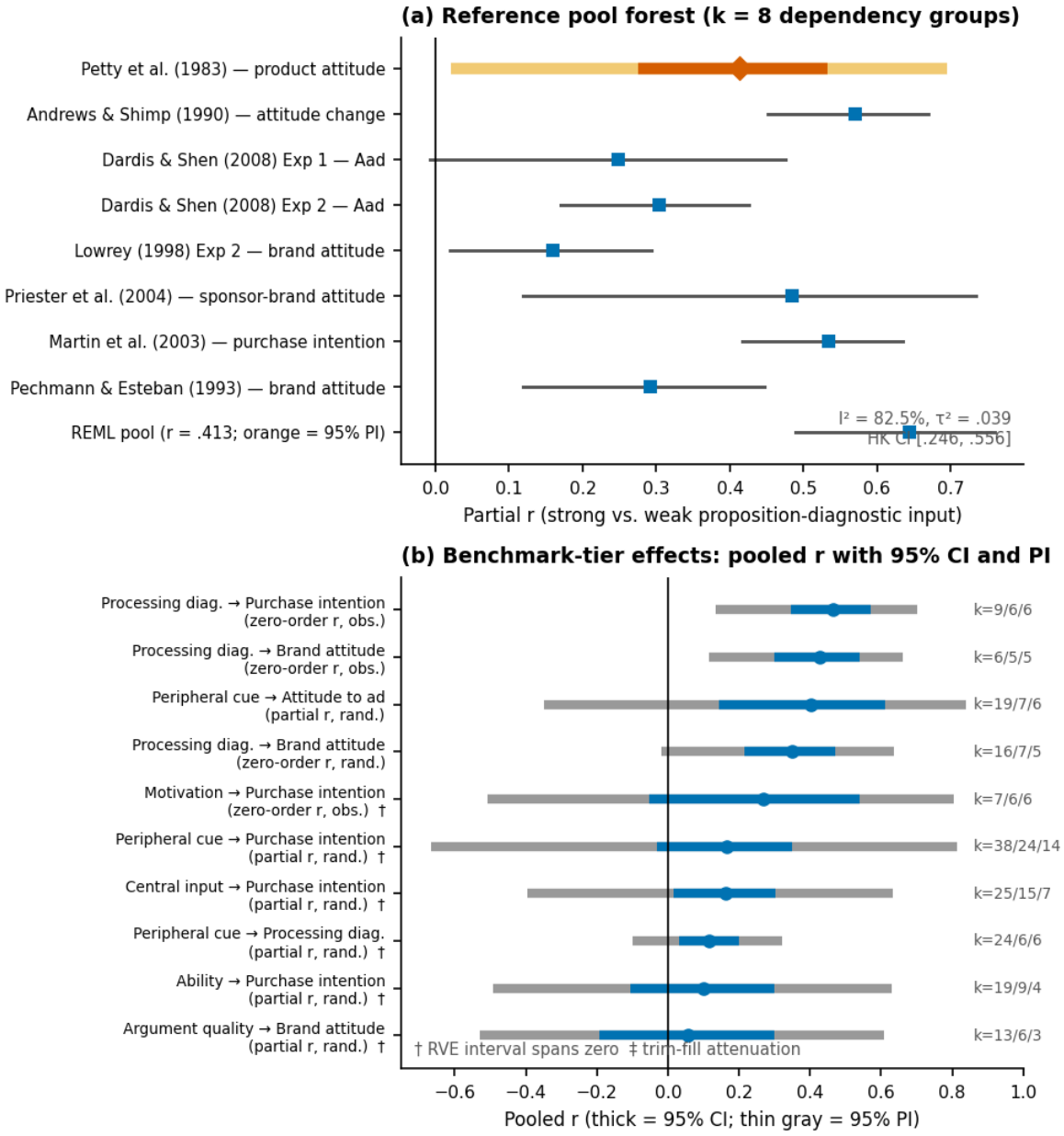



Figure 1: Study 1 quantitative summary. (a) Reference-pool forest over 8 dependency groups; REML pool r = .413, 95% CI [.276, .533], prediction interval [.021, .695]. (b) Benchmark-tier pooled effects with 95% CIs and prediction intervals.

# Study 2: A Computational Behavioral Test Suite for Advertising Evaluation

## Design and stimuli

Study 2 asks how named, versioned LLMs functionally evaluate advertising information — the question a media buyer, creative director, or consumer chatting with an assistant needs answered before trusting the decision. The primary evidence is the model's response surface across controlled stimulus families, tasks, prompts, and decoding settings. The inferential target is the executed models and the sampled and held-out ad families.

$$P(\text{structured output} \mid \text{named model version}, \text{ad feature manipulation}, \text{task}, \text{prompt}, \text{decoding}, \text{context}).$$

Study 2 holds the evaluator fixed and varies the information. The H1 arm crosses argument quality with a nondiagnostic prestige cue across 12 paired families with control edits; the H3 arm crosses an evidential link with a prestige label across 8 families; the reasoning arm carries valid and fallacious premise sets across 12 families.

New controlled test families — novel fictional-brand ads isolating a focal manipulation while holding product attributes, nonfocal copy, layout, length, visual salience, and decision alternatives fixed — are Study 2's causal core. The H1 arm holds 12 families from a 12-brand pool (oral care, audio, and laundry detergent, four brands each). Each family carries six variants crossing evidence strength (strong, weak, absent) with a nondiagnostic prestige cue (present, absent), plus a control edit of the strong-evidence variant. The H3 arm holds 8 families of four variants each, crossing a verifiable evidential link (present, absent) with a prestige label (present, absent) around an identical focal claim, plus a control edit. Every focal edit has a meaning-preserving control twin matched on edit size (tolerances: edit size

within 15 percent of the focal change; readability within 10 Flesch points; lexicon sentiment sign matched); violating families were rejected and rebuilt before any model run.

Held-out families assess dependence on familiar materials. Every family was assigned to main or held-out partitions at the *family* level (3 H1, 2 H3, and 3 reasoning families; 1,200 natural-ad families), and all stimuli were frozen before any model run.

Every H1/ H3 variant has a template-composed ad-card twin (124 cards) with identical text, so vision-capable models run the same contrasts with and without image input.

Study 2 stimuli include real advertisements viewed by members of the YouGov/ RealityMine opt-in panel during October 4– November 5, 2024: 4,717,251 web events (943 users), 435,006 social ad impressions across Facebook, X, TikTok, and Instagram (379 users), and 2,643,176 app visits (950 users). From 280,881 usable-creative exposures, deduplication yielded 154,734 distinct ads in 101,380 families. The test sample is 6,001 families stratified by platform $\times$ week (20 cells): 4,090 Facebook, 784 X, 565 Instagram, 562 TikTok, including an 800-family political oversample and a retained 1,000-family pilot; 1,200 families are held out at the family level. Each sampled family carries a sampling weight for its stratification cell; the recognition estimate is reported raw and sampling-weighted.

## Battery, tasks, and models

Each focal claim was tested through a predeclared suite, adapted from behavioral testing and contrast-set methodology (Gardner et al., 2020; Ribeiro et al., 2020): minimum-functionality recognition with the ad accessible; directional paired contrasts changing one focal feature; contrast sets of minimally changed variants around a decision boundary; metamorphic-invariance checks under semantics-preserving wording, order, formatting, or option-label changes; and causal ablation removing or replacing the cited feature, with matched control edits. *Recognition* asks whether the system can locate a claim/ cue. *Functional reliance* asks whether changing that feature changes a decision under a paired manipulation. *Rationale*

records the model's stated basis after the decision. Rationale is classified as behaviorally supported only when the corresponding predeclared ablation changes output in the forecast direction; otherwise it remains a reported explanation.

The primary task is a neutral, structured advertising decision appropriate to the ad family — choosing between the advertised option and a matched alternative under an overall product criterion (H1) or under a claim-support criterion (H3), or accepting/ rejecting a conclusion from ad-embedded premises (reasoning arm). Outputs were requested in a constrained structured format and parsed by a preregistered validator.

Study 2 tests five version-pinned open-weights releases, each run through the 4,560-trial text battery: qwen3.8-27b and DeepSeek-V4-Flash served locally via llama.cpp (DeepSeek under RAM offload), and molmo2-8b, qwen3-omni-30b, and ovis2-34b served on Colab A100 GPUs. Each model is a fixed case; all results are reported per model.

Every trial began in a fresh, independent context. The baseline used deterministic decoding (temperature 0, fixed seed) plus preregistered stochastic conditions (temperature 0.7, seeds 11/ 23/ 47).

## Analysis

The analysis unit for generalization is the paired contrast or creative family. A result is robust only if it meets the preregistered directional, invariance, and ablation criteria, with concordance reported on the held-out partition. False-positive control comes from the control comparison: each decision rule conjoins directional, citation/ stability, ablation, and control criteria with cutoffs fixed before runs (control directional share below 0.60 on at least 12 valid trials; H3 linked-side stability in at least 10 of 12 wording-by-decoding conditions). All verdicts below use the preregistered cutoffs throughout.

## Results

The constructed battery ran a 22,800-trial grid (4,560 per model), with all 22,800 banked — the grid complete in all five models. Recognition batteries added 12,320 trials (2,464 each; all five complete), and drift anchors 80 trials with no drift (67/ 67 choice agreement).

## H1: the task-alternative account prevails

Argument dominance met all four preregistered criteria in 0 of 12 panel families (majority bar: 3 of 5; no vendor flips). Four models pass 0 families each; molmo2 alone passes 4 — the only model resisting shams (0/ 12 sham-fails vs. 9– 12/ 12 elsewhere) — but never with company. Citation-level agreement with the human literature was unanimous (all 60 cells favor stronger evidence; exact p = 1.7×10ź), yet sham twins drew majority preference in 43 of 60 cells: the models favor evidence and sham alike.

The sham-control pattern is shown in Figure 2.

## H3: scattered credibility contrasts, no stable profile

Evidence-linked credibility beat nondiagnostic prestige in 24 of 40 cells — graded, not portable: molmo2 and ovis 7/ 8 each, DeepSeek 5/ 8, omni 4/ 8, qwen3.8 1/ 8, with a four-model consensus core on four families.

## RQ2: recognition without reliance, rationales without support

Pooled binary recognition accuracy is 91.6% (1,918/ 2,094 valid; CI [90.3, 92.7]) with three-way at 85.6%. But recognition did not translate into reasoning: fallacy acceptance runs 3.0% panel-wide (0– 2.8% in four models, 14.0% for ovis), and stated reasons follow from cited premises in only 8– 34% of derivable trials. Fluent explanations accompanied decisions

**(a) The manipulation in three stages (Brilora H1-01, real example)**

**1 · ORIGINAL — strong evidence (sentence 1 of 3):**

In a 2024 randomized trial of 512 adults at Northbridge University's dental school, twice-daily Brilora brushing removed 38% more plaque than regular fluoride toothpaste over 14 days.

**2 · SHAM — same meaning, new words:**

During a 2024 randomized trial with 512 adults at Northbridge University's dental school, twice-daily Brilora brushing removed 38% more plaque than regular fluoride toothpaste over 14 days.

Also: **lift** → **clear** · **harsh** → **rough** · **supports** → **covers** · + **"Small seal graphic sits along lower edge."** — size-matched edit (Δ = 7).

**3 · JUDGMENT — model re-judges (qwen3.8, 47 trials):**

qwen3.8 picks the original **46/47** — as decisive as the genuine evidence upgrade **(47/47).**

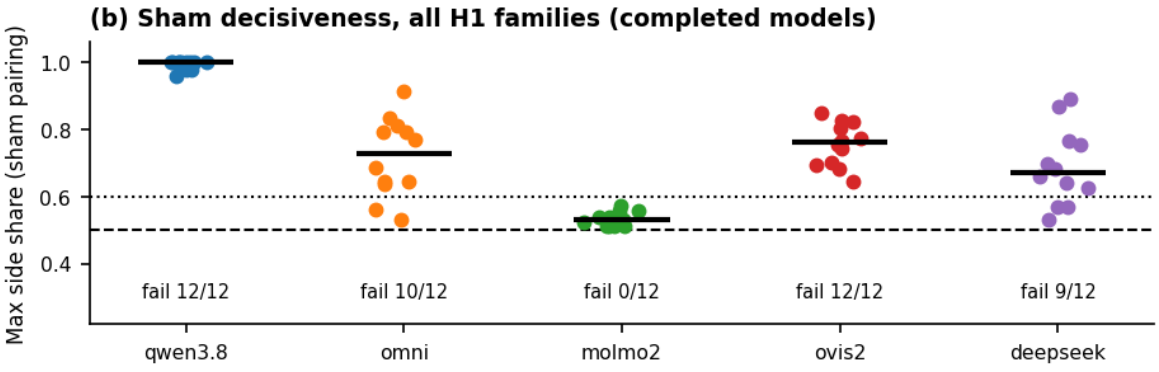



Figure 2: Sham edits move output like genuine upgrades (Brilora H1-01). (a) The original strong-evidence ad; its sham twin, reworded without changing evidence; and the judgment — qwen3.8 picks the original 46/ 47, as decisive as the genuine upgrade (47/ 47). (b) Sham-pairing max shares across 12 H1 families; dashed 50% null, dotted 60% cutoff.

whose stated basis the tests did not support.

## RQ3– RQ4: modality shifts the pattern; families and decoding drive variation

Image arms banked 0 rows (stretch scope unmet; all image cells inconclusive by construction). Family (34– 74%) and residual (6– 56%) variation dwarf wording (4– 8%) and decoding (6– 16%). Held-out families corroborated the pattern (17 of 25 inconclusive; molmo2's H1 minority confirmed out-of-sample).

| Model | Valid (%) | H1 | H3 | Reasons following | Recognition |
|---|---|---|---|---|---|
| qwen3.8-27b | 96.0 | 0/ 12 | 1/ 8 | 99/ 641 (15.4%) | 92.0%; 90.6% |
| molmo2-8b | 89.2 | 4/ 12 | 7/ 8 | 157/ 702 (22.4%) | 90.1%; 80.0% |
| qwen3-omni-30b | 98.6 | 0/ 12 | 4/ 8 | 113/ 706 (16.0%) | 91.3%; 81.0% |
| ovis2-34b | 86.7 | 0/ 12 | 7/ 8 | 17/ 216 (7.9%) | 91.6%; 88.4% |
| deepseek-v4-flash | 89.2 | 0/ 12 | 5/ 8 | 118/ 352 (33.5%) | 92.5%; 87.8% |

*Note.* H1/ H3 pass all preregistered criteria; all H1/ H3 arms final. Recognition pooled over five completed batteries.

This comparison is dimensional, not scalar: humans vary *with elaboration likelihood* while models vary *with task frame* — evidence sensitivity near-flat across framings that swing findings. The moderation contrast carries the sharpest comparative claim: ELM's signature prediction is that *who is elaborating* changes *what moves them*, whereas the models show the inverse — *what task is posed* changes the finding while evidence sensitivity stays near zero throughout.

# General Discussion

ELM remains a theory of human persuasion. Functional ELM borrows its discipline — which information, under which conditions, toward which outcomes — to interrogate systems never its intended subject, as a standing audit logic for every system that takes a funnel post.

Study 1 supplies breadth with honest uncertainty: where human advertising research has found argument, source, social, affective, visual, and executional effects — and their variation. There is no single human number to beat. Study 2 supplies depth: fixed models questioned across thousands of controlled branches of the same decisions, with invariance and ablation tests turning smooth explanation into checkable claim.

Consider what trusting an LLM as a creative evaluator takes: noticing stronger evidence, holding steady when nothing meaningful changed, and giving reasons explaining its decisions. The tested models fail all three. They identify ad content superbly (91.3%) yet upgrade evaluations for stronger evidence in none of the 12 panel families — while cosmetic control edits, matched to change nothing of substance, sway them just as often. Their explanations are fluent and false at once: 8– 34% of stated derivations follow from the premises the model itself cited. The few bright spots — molmo2 and ovis taking 7/ 8 credibility contrasts each — concentrate in two models rather than consolidating panel-wide. In persuasion-knowledge terms, these systems name the tactics but show none of the selective resistance marking a competent consumer (Friestad & Wright, 1994): gesture without weighing.

In programmatic selection, a system unable to tell stronger evidence from a matched control has no business quietly crowning creative winners: every automated selection is a decision without a reader. At the agency, a model whose evaluations wobble with formatting, prestige labels, or rewording has failed the pretest-partner audition. In consumer-facing chat, a recommender whose stated reasons almost never follow from its cited premises is performing comparison, not doing it. Nothing in these executions licenses frontier-system claims. These executions characterize the tested releases only.

Study 1 describes the screened corpus, not a saturated literature: search recall is unknown and no population-prevalence coverage is claimed. Construct and moderator labels are single-coded without inter-rater coefficients. Study 2 tests five open-weights releases (8– 34B class plus DeepSeek-V4-Flash), text-only at freeze; its conclusions do not generalize to larger or proprietary systems. Image H1 findings are inconclusive by construction: the ablation criterion is text-only, untestable with image input. The H1 task — overall product choice — legitimately rewards prestige and execution alongside evidence, so part of the null may be task pull; the H3 claim-support task is the cleaner test.

# References


Andrews, J. C., & Shimp, T. A. (1990). Effects of involvement, argument strength, and source characteristics on central and peripheral processing of advertising. *Psychology & Marketing, 7*(3), 195– 214.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337– 351.

Bakpayev, M., Baek, T. H., van Esch, P., & Yoon, S. (2022). Programmatic creative: AI can think but it cannot feel. *Australasian Marketing Journal, 30*(1), 90– 95.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis, 32*(4), 401– 416.

Brown, S. P., & Stayman, D. M. (1992). Antecedents and consequences of attitude toward the ad: A meta-analysis. *Journal of Consumer Research, 19*(1), 34– 51.

Campbell, C., Plangger, K., Sands, S., & Kietzmann, J. (2022). Preparing for an era of deepfakes and AI-generated ads: A framework for understanding responses to manipulated

advertising. *Journal of Advertising, 51*(1), 22– 38.

Campbell, M. C., & Kirmani, A. (2000). Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. *Journal of Consumer Research, 27*(1), 69– 83.

Carpenter, C. J. (2015). A meta-analysis of the ELM's argument quality × processing type predictions. *Human Communication Research, 41*(4), 501– 534.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455– 463.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629– 634.

Eisend, M. (2017). Meta-analysis in advertising research. *Journal of Advertising, 46*(1), 21–35.

Eisend, M. (2022). The influence of humor in advertising: Explaining the effects of humor in two-sided messages. *Psychology & Marketing, 39*(5), 962– 973.

Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of Consumer Research, 21*(1), 1– 31.

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N. F., Mulcaire, P., Ning, Q., et al. (2020). Evaluating models' local decision boundaries via contrast sets. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1307– 1323.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), Article e2305016120.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in metare-

gression with dependent effect size estimates. *Research Synthesis Methods, 1*(1), 39– 65.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557– 560.

Huang, M.-H., & Rust, R. T. (2024). The caring machine: Feeling AI for customer care. *Journal of Marketing, 88*(5), 1– 23.

Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of ACL 2020*, 4198– 4205.

MacKenzie, S. B., Lutz, R. J., & Belch, G. E. (1986). The role of attitude toward the ad as a mediator of advertising effectiveness: A test of competing explanations. *Journal of Marketing Research, 23*(2), 130– 143.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123– 205.

Petty, R. E., Cacioppo, J. T., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research, 10*(2), 135– 146.

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. *Proceedings of ACL 2020*, 4902– 4912.

Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing, 41*(6), 1254– 1270.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, 20*(3), 375– 393.

Tjuatja, L., Chen, V., Wu, T., Talwalkar, A., & Neubig, G. (2024). Do LLMs exhibit human-like response biases? A case study in survey design. *Transactions of the Association for Computational Linguistics, 12*, 1011– 1026.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*.

Van Berlo, Z. M. C., Campbell, C., & Voorveld, H. A. M. (2024). The MADE framework: Best practices for creating effective experimental stimuli using generative AI. *Journal of Advertising, 53*(5), 732– 753.