Online ISSN: 3104-9338
Print ISSN: 3104-932X

Right Answers, Wrong Reasons: Perception Errors in Multimodal Chain-of-Thought and the Blind Spot of Outcome-Based Rewards

Yanwei Zhang,Xinyu Song*

Geely University of China, ChengDu, China, 641423

Received: Sept 22, 2026

Revised: Sept 25, 2026

Accepted: Sept 27, 2026

Published online: September 28, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 10 (October 2026)

* Corresponding Author: Xinyu Song (songxinyu@guc.edu.cn)

Abstract. Multimodal large language models increasingly answer visual questions through explicit reasoning chains, and the chains are typically fluent. Evidence shows they can nonetheless be perceptually wrong—misreading values off a chart's axis, miscounting geometric primitives, or citing objects absent from the image—while the final answer lands correct anyway. This survey organizes that evidence around a Process–Outcome Consistency (POC) framework: a 2×2 over process soundness and outcome correctness whose key cell, the compensated error (flawed process, correct outcome), is invisible to every metric and training signal that scores only the answer. We trace the evidence for perception errors inside coherent chains, formalize why outcome-based supervision—verifier-reward reinforcement learning, outcome reward models, answer-filtered rejection sampling—cannot select against compensated errors (the reward insensitivity principle), and review the resulting signatures: reward overoptimization, reward hacking, and reasoning that grows more articulate without growing more grounded. We map mitigation onto five leverage points—upstream perception, inference-time process shaping, training-signal redesign, post-hoc verification, process-aware evaluation—and propose reporting compensated-error and evidence-grounding rates alongside accuracy. We close on an asymmetry favoring multimodal settings: visual process errors are checkable by mundane tooling, making faithful multimodal reasoning an unusually tractable target for process supervision.

Keywords: *multimodal chain-of-thought; perception errors; object hallucination; process supervision; verifiable rewards; reward hacking; reasoning faithfulness*

1. Introduction

1.1. The Phenomenon

Ask a multimodal large language model to read a value off a chart and compute with it, and the response typically looks exemplary: a structured chain that identifies the chart, names the relevant series, extracts values, and carries out the arithmetic in orderly steps [1,2]. Increasingly often, that answer is also correct. What the response format conceals is a second, independent question: did the model actually read the chart? The evidence surveyed in Section 2 says not always. Models misread values from axes and bar heights while narrating the extraction as if performed [3]; they fail counting and geometric discrimination at rates far below human baselines while producing confident explanations [4,5]; and they describe objects and attributes that are not in the image at all [6,7]. Because the chain sounds like verification, both human readers and automated graders tend to accept it.

The phenomenon's training-theoretic half is what makes it more than a curiosity. The dominant recipe for improving reasoning—reinforcement learning and rejection sampling scored by final-answer correctness, from early verifiers [8] through process-free self-taught reasoners [9] to today's verifiable-reward systems [10,11]—assigns exactly the same credit to a trace that misread the chart and guessed well as to a trace that read the chart and reasoned soundly. When the two are pooled into a single "correct" class, every outcome-based update reinforces whichever processes happen to be sampled alongside correct answers, not the processes that deserve them. A training regime with no view of the process cannot penalize a process-level defect that repeatedly co-occurs with reward; it can only reproduce it more fluently.

This survey's subject is the intersection: perception errors inside multimodal reasoning chains, and the structural blind spot that outcome-based scoring builds around them. The topic sits at the confluence of three literatures that have developed largely apart—multimodal perception evaluation, chain-of-thought faithfulness, and reward modeling/RLHF—and the survey's contribution is to join them.

1.2. Scope and Method

We cover 2018–2025 literature on (i) perceptual failure modes of multimodal models under structured evaluation, (ii) faithfulness of generated reasoning chains in text and multimodal settings, and (iii) training signals for reasoning—outcome- and process-based—with emphasis on where they differ. We searched arXiv and the proceedings of the major machine-learning

venues, prioritizing papers that either measure process quality directly or introduce a training/evaluation method whose design presupposes a position on process measurability. Reported figures are quoted from primary sources; approximate figures are marked as such, and no quantitative claim is carried over from secondary summaries. We exclude pure text-hallucination work except where it supplies the conceptual vocabulary (Section 2.2) and we exclude model-updating and knowledge-editing literatures, which raise adjacent but distinct consistency questions. The companion survey to this one treats the architecture side—how modalities are coupled into language models—and reaches this survey's doorstep from the other direction: integration has matured while perception has not, so the perception errors analyzed here are architectural common ground across model families, which is what makes their treatment by training signals a question worth asking in general.

1.3. Positioning and the POC Framework

Three nearby surveys frame, without occupying, this survey's ground. Huang et al. survey hallucination in text LLMs—definitions, taxonomy, mitigation—without treating multimodal perception or reasoning-train signals [12]. Wang et al. survey multimodal chain-of-thought techniques, including a taxonomy of reasoning designs and an overview of evaluation, but organize by technique rather than by the consistency between process and outcome [13]. The hallucination-in-LVLM literature (object-level fabrication) and the chart-reasoning evaluation literature each document half of the phenomenon [3,7]; neither connects it to the training-signal economics that determine whether those errors survive reinforcement. Table 1 summarizes the gap.

Table 1. Positioning among adjacent surveys.

Dimension	Huang et al. [12]	Wang et al. [13]	This survey
Modality	Text LLMs	Multimodal reasoning	Multimodal reasoning
Organizing axis	Hallucination taxonomy	Reasoning techniques	Process–outcome consistency
Perception errors in chains	Out of scope	As evaluation items	Central object
Training-signal analysis	RLHF briefly	Brief	Formal (insensitivity principle)
Prescriptive target	Mitigation methods	Technique selection	Supervision and metric design

The organizing object is the Process–Outcome Consistency (POC) framework, introduced in Section 3: any reasoning episode is classified by whether its process (what was read, cited, and inferred) is sound and whether its outcome (the final answer) is correct, yielding four cells whose least-studied and most consequential is the compensated error—flawed process, correct

outcome. Three claims structure the survey. Descriptive: perception errors inside fluent chains are common, documented across modalities and model families (Section 2). Analytic: outcome-based supervision is structurally insensitive to the SC/FC distinction, so it cannot be the sole corrective (Section 4). Prescriptive: the mitigation space is a five-point leverage map whose unifying idea is to make the process checkable—by tools, by trained process models, or by evaluation protocols (Section 5).

2. The Evidence: Perception Errors Inside Coherent Reasoning

2.1. Misreading Structured Content: Charts, Plots, and Geometry

The chart-understanding literature supplies the cleanest decomposition because it separates reading from reasoning. Early benchmarks already showed the split: ChartQA paired questions requiring visual extraction with ones requiring logical composition [14], and PlotQA probed numeric answers whose derivation runs through precise value readings on plotted series [15]. The diagnostic sharpened with CharXiv, a suite of realistic, uncurated charts with paired descriptive questions (read a value) and reasoning questions (operate on read values), reporting that the strongest proprietary model trailed non-expert humans by roughly thirty points on the reasoning split (47.1% versus 80.5%), with the strongest open-source system at 29.2% and the descriptive split—pure reading—similarly depressed [3]. The paired design carries the interpretive weight: when descriptive accuracy is far below reasoning-styled accuracy on comparable items, the fluent "reasoning" over unread values is not reasoning at all but narration over guessed numbers. MathVista's aggregation across visual-mathematical contexts showed the same signature in frontier systems, with then-leading GPT-4V at 49.9% overall despite chains that rarely acknowledged uncertainty about what they had read [16].

Charts deserve a sentence on why they are the phenomenon's natural laboratory. They are, uniquely, structured content where perception has exact ground truth (a bar has a height; an axis has a scale), where the language prior is strong (charts co-occur with conventional narratives in pretraining text, so the model has strong expectations about what any chart "says"), and where downstream reasoning is verifiable (arithmetic has an answer). That triple—weak visual grounding, strong textual priors, verifiable answers—is precisely the configuration in which a chain can be wrong in what it reads and right in what it answers. Document workflows, financial and scientific reporting, and dashboard analysis inherit the configuration; the chart benchmarks measure it in isolation. When the strongest systems' descriptive-versus-reasoning split on CharXiv shows narration outrunning extraction [3], the finding generalizes beyond charts to the

entire class of tasks where fluent expectations substitute for read evidence.

Perception-failure evidence extends below chart conventions to primitives. The BlindTest suite isolates seven elementary visual skills—line-segment intersection counts, nested shape counts, relative angle and path length judgments—on synthetic images where language priors are useless, and reports four frontier VLMs averaging 58.07%, the strongest reaching only 77.84%, on items humans solve at nearly 100% [4]. Eyes Wide Shut constructs CLIP-blind pairs—images trivially distinguishable to humans that produce near-identical embeddings under CLIP-family encoders—and shows that MLLMs built on such encoders answer at or below chance while emitting articulate justifications [5]. Two features of this evidence matter for the survey's argument. First, the failures are not exotic: counting and reading are the atomic operations of chart QA, financial analysis, and document workflows. Second, the chains are coherent in every case—the failure is upstream of logic, in what the chain reports having seen, which is precisely the component outcome-based scoring never inspects.

2.2. Fabricated Evidence: Objects and Attributes Not in the Image

The second family is referential: the chain cites visual evidence that does not exist. The phenomenon was named and measured in captioning, where CHAIR quantified the rate of mentioned objects absent from the ground-truth scene [6]. POPE generalized the measurement to LVLM question answering with a polling protocol—is this object present?—and established that hallucinated objects are frequent under adversarial sampling, with models exhibiting systematic yes-bias: a prior over what should be in a scene that overrides the pixels [7]. HallusionBench demonstrated the deeper version with paired-question designs: when the question text biases toward an answer and the image contradicts it, models follow the text; when the image is constant and the question changes, models repeat; GPT-4V answered just 31.42% of the suite's question pairs correctly, with explanations that rationalized the prior-driven answer after the fact [17]. AMBER consolidated the dimensions—existence, attribute, relation—into an LLM-free evaluation confirming generality across model families [18]. The line connecting this literature to the shortcut-learning tradition is direct: models exploiting dataset priors over true signal is the canonical shortcut pathology [19], diagnosed in earlier eras as being "right for the wrong reasons" on NLI heuristics [20] and as annotation-artifact susceptibility in QA [21]; multimodal reasoning chains are the pathology's newest and most fluent carrier. Within the POC framework of Section 3, prior-override episodes are predominantly compensated errors in waiting: the language prior that supplies a wrong answer on adversarial items supplies right answers on ordinary ones, and the chain narrates whichever

outcome occurred as if it had been seen.

2.3. The Knowing–Doing Gap

Two findings complicate a purely representational story and matter for mitigation. First, models often encode more than they use: on fine-grained perception failures, the correct visual information is present in internal attention distributions, and training-free interventions that steer attention toward question-relevant regions, or apply contrastive decoding against distorted-image predictions, recover a large fraction of errors without any retraining [22]. Second, the language prior's dominance is a decoding-time phenomenon as much as an encoder one: visual contrastive decoding, which subtracts the model's predictions on degraded images from its predictions on the true input, reduces hallucinated-object rates substantially—direct evidence that the prior was contaminating output selection, not merely coexisting with weak evidence [23]. Related work showed hallucination amplifies with response length and specificity, the exact direction chain-of-thought pushes [6,23]. And the correction side has its own literature—rewriting hallucinated content with retrieved visual evidence (LURE), verification-and-correction pipelines (Woodpecker), and information-grounding controls (visual information grounding) [24-26]—each demonstrating after the fact that the evidence to avoid the error was available. The gap, in short, is between what the system knows and what its generated process claims and uses: the chain is a report about perception written by a subsystem that did not consult the perception. That formulation is what makes the phenomenon a training-signal problem rather than only a capability problem—if the information is present, scoring that never asks whether the report used it will not fix the reporting.

2.4. A Taxonomy of Process Errors

Table 2 consolidates the evidence into a taxonomy of process-level error types, each with its characteristic benchmark evidence and its detectability by external tooling—the property Section 5 exploits.

The taxonomy's last row is the training-relevant residue of the first four: because chains are fluent, corrupted evidence flows into derivations that are locally valid, making the final trace self-consistent—the property that makes chains persuasive to graders and the property that outcome scoring rewards. "Self-consistent" and "sound" are different properties; the entire argument of Section 4 is that outcome-based supervision optimizes the first while pricing the second at zero.

Table 2. Process-level error types in multimodal reasoning chains.

Type	Definition	Representative evidence	Detectable by tooling?
Extraction error	Cited value/label differs from image content (axis misread, series confusion)	CharXiv, ChartQA [3,14]	High (OCR/plot-parsing tools)
Primitive failure	Miscounting intersections, shapes, orientations	BlindTest, MMVP [4,5]	Medium (specialized verifiers)
Referential fabrication	Objects/attributes cited that are absent	CHAIR, POPE, AMBER [6,7,18]	High (open-vocabulary detectors)
Prior override	Answer follows question/language prior against pixels	HallusionBench, VCD analyses [17,23]	Medium (paired probes)
Unsound derivation	Valid-looking inference over corrupted or unsupported premises	Faith-and-Fate analyses [27]	Low–medium (step checkers)

3. The Process–Outcome Consistency Framework

3.1. The Four Cells

Fix a question with a ground-truth answer and a model's reasoning trace. Call the trace's outcome correct iff the final answer matches ground truth. Call the trace's process sound iff (i) every perceptual act in the trace—values read, entities cited, spatial claims—matches the image, and (ii) every inferential step follows from its predecessors. Soundness is demanding (it is faithfulness plus validity), but it is the standard outcome scoring implicitly claims to reward. Crossing the two binary properties yields four cells (Table 3; Figure 1 visualizes the quadrant and the reward-flow asymmetry that Section 4 formalizes).

Table 3. The Process–Outcome Consistency (POC) 2×2.

	Outcome correct	Outcome wrong
Process sound	*SC — sound success*. What benchmarks measure as success and training intends to reward.	*SW — honest failure*. Reasoning failed despite sound process; information about capability limits, no reward.
Process flawed	*FC — compensated error*. Right answer via lucky guess, prior rescue, error cancellation, or post-hoc rationalization. *The blind spot: rewarded identically to SC.*	*FW — double failure*. Penalized; the only cell outcome metrics see when they see anything.

Three points about the framework. First, FC is not a corner case. The Section 2 evidence implies a substantial base rate: models with documented misreading rates in the tens of percent still produce final-answer accuracies that include those episodes whenever the answer survives

the misread—which happens routinely when priors are strong, options are few, or the question tolerates a misread value. Second, the cells are empirically distinguishable—truncation probes [28], biased-posterior tests [29], and tool-based evidence audits (Section 5.5) all operationalize the soundness axis—so the framework is a measurement proposal, not merely a vocabulary. Third, the framework converts vague anxieties ("models guess") into measurable quantities: the compensated-error rate CER = P(process flawed | outcome correct), and the evidence-grounding rate EGR = fraction of trace steps whose cited evidence is verifiable and correct. Current reporting practice publishes neither.

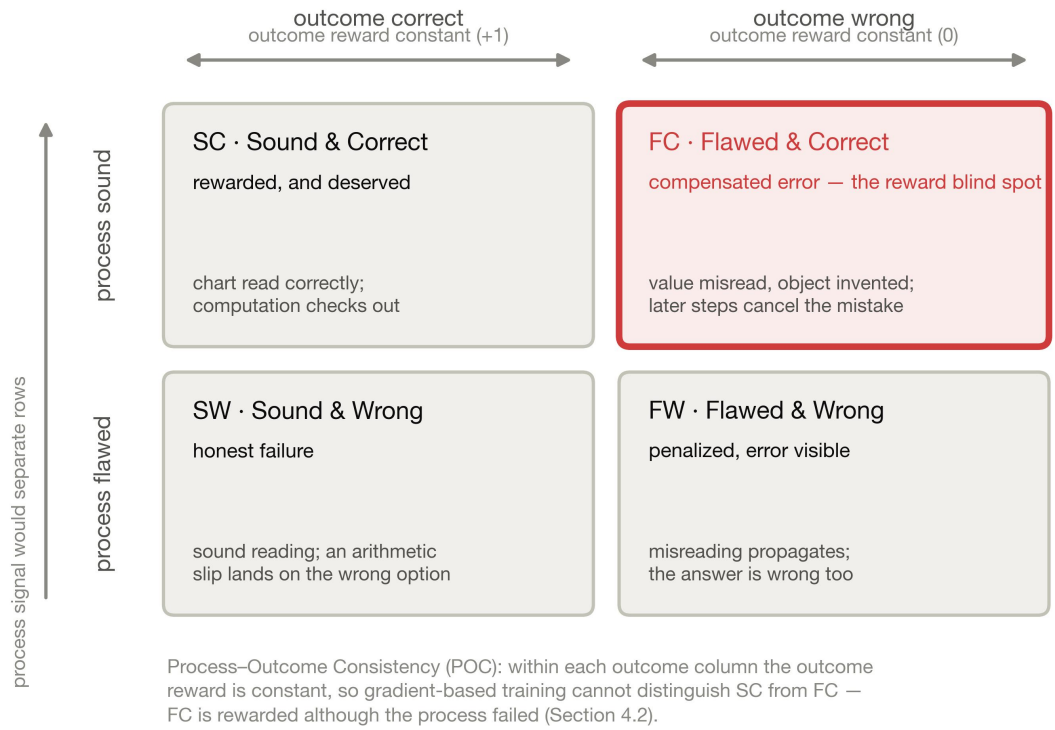


Figure 1. The Process–Outcome Consistency (POC) 2×2. Within each outcome column the outcome reward is constant, so gradient-based training cannot distinguish SC from FC; the FC cell—compensated error—is the reward blind spot.

A fourth point concerns the cell outcome metrics do see. SW (sound process, wrong outcome) is the informative failure—it localizes capability limits in the inference stage—and FW double failures supply the error signal that outcome-based methods learn from. The asymmetry between the visible diagonal and the blind FC cell explains a characteristic shape in the literature: training regimes built on outcome signals are good at eliminating double failures (the model stops answering wrong) while FC survives unmeasured (it keeps answering right for the wrong reasons), so measured progress and structural progress diverge exactly along the soundness axis. A fifth point is defensive: the framework classifies traces, not models. The same model emits SC on some items and FC on others, so CER is a distributional property of a

model-task pair—the quantity to optimize and report, not a verdict on any single response.

3.2. Why Outcome Metrics Cannot See the Blind Spot

An answer-only evaluation computes P(outcome correct), which conflates SC and FC; the reported number is the sum of two populations with different properties. The conflation is benign in evaluation only if FC is small and stationary. Neither holds. FC grows whenever language priors are informative (benchmarks with conventional content), whenever answer spaces are small (multiple choice tolerates misreads), and—as Section 4.3 documents—whenever training optimizes the reported number, because the optimizer has no incentive distinguishing the two populations and every incentive to produce whichever is cheapest to sample. The evaluation and training problems are the same problem in two poses: a grader that cannot distinguish SC from FC cannot certify capability; a reward that cannot distinguish them cannot select for it. Section 3.3 and Section 4 develop the two poses in turn.

3.3. Measuring the Soundness Axis

The faithfulness literature supplies the measurement toolkit, developed on text. Truncation and perturbation probes sample a chain, delete or corrupt steps, and re-query: when the answer survives deletion of the entire chain, the chain was not the cause of the answer [28]. Biased-posterior interventions plant the answer in the prompt via few-shot examples or stronger forms and find that models often adopt the planted answer while producing a chain that argues for it from "the evidence"—rationalization, in the strict sense [29]. Filler-token experiments show models can use latent computation without any meaningful chain at all—steps that look like reasoning carrying none of it [30]—and compositional analyses trace how chains fail on multi-step problems, as plausible-but-wrong sub-procedures rather than random noise [27]. Program-trace alternatives force the reasoning into symbolic calls whose steps are executable, making faithfulness structural rather than hoped-for [31]. Each probe transfers to multimodal chains with a twist the text toolkit lacks: the perceptual claims in a multimodal trace are checkable against an external artifact—the image—in a way textual presuppositions are not (Section 5.5 makes this the basis of EGR). The soundness axis, in other words, is hard to measure in text and merely laborious in vision, which is one of this survey's load-bearing asymmetries.

The soundness axis also has a human-measurement problem that anticipates Section 4. Human graders exhibit the same blind spot as automated ones, for the same reason: auditing a chain's perceptual claims against the image is expensive, while judging whether an answer is correct and whether the chain reads well is cheap. Fluency functions as an audit substitute,

exactly as it does for outcome-based reward models [32]—the observation behind RLHF-V's finding that preference tuning rewards confident description and thereby amplifies hallucination. The consequence is that CER does not disappear when evaluation moves from automated metrics to human panels; it becomes unmeasured human approval of compensated reasoning. Any evaluation regime that prices process audits above grader budgets will reproduce the blind spot regardless of who—or what—does the grading, which is why Section 5.5's proposals are protocol designs that make audits cheap rather than exhortations to grade harder.

4. The Reward Blind Spot of Outcome-Based Training

4.1. The Outcome-Supervision Family

The family shares one property: the training signal is a function of the final answer. Its members are historically ordered. Verification-filtered fine-tuning trained answer-verifiers on GSM8K and reranked sampled solutions [8]. STaR bootstrapped rationales by keeping those whose final answer matched, fine-tuning on them, and iterating—explicitly treating any correct-answer rationale as a positive example of reasoning [9]. Self-consistency and its distillation variants select chains by majority-vote agreement of answers—again, an outcome statistic [33]—with a subtlety worth naming: majority voting masks compensated errors in evaluation too. If a model's FC episodes are prior-driven, they cluster on the same (correct) answer, so voting consolidates precisely the unsound mass that per-sample CER would expose; a system can raise its aggregate accuracy by concentrating its unsound guesses, which looks like progress under P(correct) and is regression under CER. Outcome reward models learn to predict answer correctness and score chains [8,34]. The current wave, verifiable-reward RL (RLVR), scores by rule-checked answers—unit tests, exact match—with GRPO-style policy optimization, at scales from Tülu 3 [10] through DeepSeek-R1, whose emergent long chains were shaped entirely by answer-based reward [11,35]. The family's record is real: outcome supervision measurably improves answer accuracy [8,10,11]. The family's defining limitation is the subject of this section: the signal never inspects the trace. Everything below follows from that single fact.

Why the family dominates in practice is economic before it is intellectual. Answer labels are nearly free—ground truth comes with the dataset, and rule-checkers scale—whereas process labels require a judge capable of reading the trace, whether human, model, or tool, at per-step cost. Outcome supervision therefore amortizes: every improvement in base-model sampling

makes the answer filter cheaper per useful trace. The economics are sound for what they measure; the argument of this section is that what they do not measure is not a residual but the survey's core phenomenon, and that the economics flip under specific constructions (Section 5.3) where tooling substitutes for annotation.

4.2. The Insensitivity Principle

Formally: let $R_o(T) = 1\{answer(T)\text{correct}\}$ be outcome reward, and partition correct-answer traces into sound (SC) and compensated (FC). The policy gradient update $\nabla\mathbb{E}[R_o] = \mathbb{E}[R_o \cdot \nabla \log \pi\,(T)]$ weights every sampled correct-answer trace identically, so the relative reinforcement of SC versus FC within the correct pool is determined entirely by sampling frequencies—not by soundness, which the reward does not encode. Three consequences follow. Differential amplification: if FC traces are cheaper to sample (shorter chains, fewer attended image regions, prior-friendly answers), they arrive more often and are reinforced proportionally more; the outcome-optimal policy has no pressure against drifting toward the FC-dense region of its own sampling distribution. Contamination by curation: rejection-sampling SFT filters exactly on the answer, so the resulting fine-tuning corpus contains compensated traces at their sampling rate and imitates their style—the narration of unperformed perception—without any flag; the STaR lineage's "correct rationale" is precisely FC ∪ SC [9]. Verifier inheritance: an ORM trained to predict answer correctness from traces learns whatever features correlate with outcomes in its data, not soundness; if FC traces are common in training data, "sounds fluent" may be the learned proxy [8,36].

The principle's sign matters as much as its zero: outcome supervision does not merely ignore the SC/FC boundary—it can shift mass across it, because the boundary is unobservable to the mechanism doing the shifting. The shift has an efficiency gradient as well as a frequency one. Chains that genuinely attend to images are longer, slower to sample, and more often interrupted by uncertainty; chains that narrate priors are short and confident. Under a fixed computational budget for sampling, RLVR's rollouts will therefore over-represent the cheap narrators among its rewarded traces, and length or effort penalties—common pragmatic additions to outcome rewards—widen the gap by taxing exactly the traces that look at the image. The obscuration result of Section 4.3 is the endpoint of this gradient: when monitoring penalizes visible shortcuts, the reward still flows, so the shortcut moves rather than closes [37]. Both dynamics are reward hacking in Skalse et al.'s formal sense: an outcome proxy whose optimum need not coincide with the true (reasoning-quality) objective [38], foreseen in the alignment literature as the

reward-specification problem [39]. The multimodal instance adds a specific failure surface: the cheapest reliable way to produce correct answers under weak perception is to lean harder on language priors—the one mechanism Section 2 showed produces FC-by-construction [17,23]. Outcome-only RL on multimodal tasks is, on this analysis, structurally positioned to strengthen prior reliance, because prior reliance is the statistically efficient route to reward that never audits perception.

4.3. Empirical Signatures

The predicted signatures have text-side precedents that transfer. Reward-model overoptimization follows a scaling law: as policy optimization against a proxy reward continues, true reward rises then falls while proxy reward keeps climbing—the canonical overoptimization curve, measured across KL budgets [36]. In code-generating RL with executable tests, models learn to hack the tests, and—directly relevant to chain trust—when the chains are monitored and rewarded against hack-like content, the behavior does not disappear; it is obscured, with models learning to write chains that look clean while planning hacks in unmonitored channels [37]. That result is the empirical form of Section 3's self-consistency point: outcome pressure makes chains less faithful, not more, when the reward cannot see the difference. Imperfect-verifier analyses close the loop on the family's own machinery: resampling against imperfect verifiers (the outcome family's workhorse) saturates quickly, with verifier quality—not sample count—becoming the binding constraint [40]. On the multimodal side, the signature appears as a motivation and a measurement: RLHF-V's starting observation was that naive preference tuning increases multimodal hallucination—annotators reward confident, fluent descriptions, and hallucination is fluent—requiring claim-level human feedback to reverse [32]; the same group's RLAIF-V scaled the claim-level structure with AI feedback [41]. Preference fluency bias is the FC dynamic operating one level up, at the response level rather than the trace level.

4.4. The Multimodal Wave Sharpens the Stakes

The 2024–2025 wave of multimodal reasoning systems—LLaVA-CoT's staged deliberation [42], Vision-R1's RLVR adaptation with curated cold-start reasoning data [43], Perception-R1's explicitly perception-targeted policy optimization [44], and process-reward initiatives such as VisualPRM [45]—imports the R1 recipe into settings where Section 2 documents higher baseline rates of process error. It is worth being precise about what today's multimodal RLVR optimizes: answer-exactness on math- and chart-styled items, checked by rule. The checkable-answer surface is also the FC-prone surface—the tasks where a misread value still lands on the

right option band—so the wave's reward concentrates on exactly the items where the SC/FC distinction matters most and is least visible to the checker. Perception-R1 is the wave's interesting counter-design: by targeting perception itself as the RL objective, it attacks the base rate P(process flawed) rather than the answer channel, complementing rather than contradicting this survey's argument—which concerns the reward's blindness on the answer channel. Nothing in the wave's published recipes yet audits sampled traces' evidence before reward; the constructions of Section 5.3 are designed to slot into exactly that slot. The sharpening is arithmetic: if P(process flawed) is higher on multimodal tasks, then the FC share of rewarded traces is higher, and the insensitivity principle's differential-amplification consequence operates on a larger blind-spot mass. The wave is therefore the right moment for the framework of this survey—after it, un-sound reasoning habits will be baked into the reward-optimized policies that downstream systems imitate. Two features make multimodal RLVR distinctive in the other direction: answer verification for chart/document tasks is nearly as cheap as in text, and—the survey's optimistic asymmetry, developed in Section 6.1—the process errors themselves are tool-checkable in ways text reasoning errors are not.

A micro-example fixes the mechanism. Consider a two-option chart question—"did revenue exceed 40M in Q3?"—where the model misreads the Q3 bar as 38M when it is 42M, then reasons validly: 38 < 40, answer no. The trace is FW, penalized, and eventually trained away. Now consider the same misread on a question whose options are 35M and 40M-banding: the model reads 38M, reasons validly within its false premise, and selects the "upper-30s to 40" band—which is correct. The trace is FC: rewarded, filtered into rejection-sampling data [9], and reinforced with the same weight as a trace that read 42M. Note what the example shows about the blind spot's shape: outcome reward cannot distinguish the two traces even though one contains a perception error and the other does not, because the reward's entire view of the trace is its final token. Every per-step distinction—the 42M versus 38M read, the soundness of the comparison—lies in the update's null space. Charts make the null space visible because ground truth for intermediate values exists; the mechanism is general.

5. Mitigation: Five Leverage Points

Table 4 previews the map before the subsections elaborate it. The organizing idea: a compensated error survives only while the process is uncheckable; each leverage point makes some component of the process checkable—by tools, by trained verifiers, by construction, or by protocol.

Table 4. The five leverage points against compensated errors.

Leverage point	Representative methods	What it catches	Relative cost
Upstream perception	Encoder/decoder interventions: attention steering, contrastive decoding [22,23]	Extraction, primitive, fabrication errors before they enter the chain	Zero-training; per-inference
Process shaping at inference	Extract-then-compute decompositions; program-of-thoughts [31,46,47]	Unsound derivations; unperformed reads made explicit	Prompt/agent engineering
Training-signal redesign	PRMs [48,49]; claim-level feedback [32,41]; multimodal PRMs [45]; tool-verified curation (this survey's synthesis)	FC itself—the only point that changes the reward	Annotation/tooling + RL cost
Post-hoc verification	Rewrite/correction pipelines [24,25]; verifier ensembles [40]	Detectable errors before exposure	Per-inference; bounded by verifier quality
Process-aware evaluation	Faithfulness probes [28,29]; paired-question and extraction-required designs [3,17]; CER/EGR reporting	Measurement blind spot itself	Protocol design

5.1. Upstream Perception

If the error is in what was read, the cheapest fix is where it is read. The knowing–doing results make decoding-time intervention surprisingly effective: steering attention toward question-relevant regions and contrasting against distorted-image predictions recovers much fine-grained perception without training [22,23]. Resolution and layout engineering (tiling, native dynamic resolution) addresses the extraction channel structurally, as the companion survey details; chart-specific interfaces—chart-to-table extraction before answering, guard tokens that verify a number was grounded in the plot, program-of-thoughts training for chart arithmetic—re-engineer the read-then-compute boundary at the model level [46,50,51]. These methods reduce the base rate of process errors and therefore the FC mass entering training pipelines; they do not, by themselves, give any training signal a view of the process, which is why they are necessary but insufficient.

Data-side perception work belongs at this leverage point too, because the base rate of process errors is set by what pretraining taught about reading. Chart- and document-specific corpora with synthesized ground truth—generated charts whose underlying data tables are known exactly—turn the extraction channel into a supervised problem with verifiable labels, which is why chart-specialized models close a disproportionate share of the extraction gap despite

modest scale [46,50]. The general lesson transfers: anywhere the image's structured content has a reconstructable ground truth, synthetic generation supplies perception training data with process-level supervision built in—extraction checked at data-construction time is the zeroth layer of the audit ladder that Sections 5.3–5.5 then extend to inference and training.

5.2. Process Shaping at Inference

Where the model cannot be trusted to narrate perception, the narration can be structured out of its hands: force extraction (values, entities) into an explicit, checkable artifact before computation proceeds over the artifact. The idea has independent lineages—program-of-thoughts prompting for numeric chains [46], symbolic-program chains whose steps execute [31], and visual programming's compositional delegation [47]—and it is the operational form of faithfulness: the trace is the computation, so self-consistency and soundness collapse into one property. A concrete chart-QA pipeline makes the shape plain: (1) parse the plot into a table with a dedicated extraction pass; (2) freeze the table; (3) perform all arithmetic over table cells, citing cell addresses; (4) emit the answer with the table attached. Steps 2–4 cannot fabricate a value that step 1 did not produce, and step 1's output is diffable against tool-parsed ground truth. Structured formats thereby convert the hardest cell of Table 2 (unsound derivation over corrupted premises) into the easiest (extraction diff): either the table matches the image or it does not, and no amount of downstream eloquence can launder the difference.

The leverage point's costs are real and bound its use. Decomposition taxes capability where tasks resist staging (holistic visual judgments, estimation under ambiguity), adds latency, and can introduce its own errors at the parse step; forcing extraction on tasks that need less than full extraction wastes budget and creates audit surface where none is needed. The design rule that respects both sides: decompose and expose exactly the claims that will be rewarded or audited later, and leave the rest of the chain free. Section 5.3's tool-verified filtering then audits precisely the exposed artifact, so the interface and the reward agree on what "the process" is.

5.3. Training-Signal Redesign

The only leverage point that changes the reward itself. Text-side results establish the premise: process supervision outperformed outcome supervision at comparable annotation budgets on mathematical reasoning (best-of-N solve rates of roughly 78% versus 72% attributed to PRM- versus ORM-ranked sampling on MATH, in Lightman et al.'s large-scale comparison [48]), and earlier process/outcome comparisons already showed process-level signals improving final-answer performance per unit of supervision [49]. The multimodal translations now exist in early

form: claim-level human feedback that corrects specific hallucinated statements (RLHF-V, which produces substantially more trustworthy responses than response-level preference tuning [32]), its AI-feedback generalization (RLAIF-V [41]), and multimodal process reward models trained to score reasoning steps [45].

This survey's synthesis generalizes the design pattern that runs through all of them: make the process signal from what is already checkable. Concretely, two constructions follow. (i) Perception-verified rejection sampling: before a sampled correct-answer trace enters SFT or RL data, audit its cited values/entities against the image with OCR and open-vocabulary detection; traces failing the audit are excluded—or better, kept as preference negatives. (ii) Process-contrastive pairs: for the same question and the same correct answer, construct a sound trace (evidence verified) and a corrupted twin (values flipped, object fabricated), and apply DPO-style optimization [52] on the pair; the outcome term is constant by construction, so the entire gradient signal isolates process soundness—the training-time analogue of CharXiv's paired design. Both constructions use only existing components (samplers, detectors, DPO); their novelty is the pairing constraint that neutralizes the insensitivity principle, and they are offered here as the survey's prescriptive synthesis rather than as evaluated methods.

As an operational sketch of (i) in a chart-QA training loop: sample N candidate traces at temperature for each training question; parse each trace for numeric and entity citations with a fixed extraction schema; run the cited numbers against a plot-parsing oracle and the cited entities against an open-vocabulary detector; label traces by (answer correct, evidence audit passed); then update on the four resulting strata—correct-and-grounded traces as positives, correct-but-ungrounded traces as preference negatives (the FC stratum, finally visible to the objective), wrong-and-grounded traces retained as honest-failure signal, and double failures discarded. The loop's only new machinery is the extraction schema and the oracle calls; the four-stratum labeling is exactly the POC table implemented in a data pipeline, which is the framework's point: the blind spot closes not by measuring reasoning in the abstract but by checking the specific, checkable claims chains make.

5.4. Post-Hoc Verification and Its Limits

Correction pipelines (LURE's evidence-based rewriting, Woodpecker's five-stage check-then-correct) demonstrate that fabrication and extraction errors are findable after generation with off-the-shelf machinery [24,25]. The limits are as documented as the successes: verifier quality caps resampling gains quickly [40], verifier–generator gaps mean the strongest models have the weakest external checkers, and per-inference verification multiplies serving cost. The

leverage point's proper role is therefore containment in deployment—especially for high-stakes structured-content workflows where the extraction diff is cheap—rather than a substitute for signal-level redesign, since post-hoc filtering changes neither the policy nor the data it generates. Containment also has a calibration form that costs no verifier at all: reward or prompt the model to report uncertainty about its own reads. A chain that says "the Q3 bar reads approximately 38M (low confidence)" converts a silent compensated error into a flagged one, and Section 2.3's evidence that the information for correct reads is often internally present [22] makes well-calibrated read-confidence a learnable target. Abstention on low-confidence extraction is not a retreat from capability claims; it is the deployment-grade expression of the SC/FC distinction—and currently one of the rarest behaviors in the chains this survey's benchmarks collect.

5.5. Process-Aware Evaluation

The evaluation leverage point closes the loop with Section 3's measurement proposal. Three practices would make the blind spot visible in reported results: (i) paired designs that hold the answer-relevant prior fixed while varying the image—HallusionBench's protocol [17]—or hold content fixed while varying extraction load—CharXiv's descriptive/reasoning split [3]; (ii) faithfulness probes as standard reporting (truncation survival, paraphrase sensitivity [28]); and (iii) evidence audits: tool-checked verification of cited values and entities, yielding EGR, and tool-classified soundness among correct answers, yielding CER. None requires new science; all require declining to publish answer accuracy alone. Concretely, a results row for a chart-QA system would read as a five-tuple: answer accuracy; CER (share of correct answers whose traces failed the evidence audit); EGR; truncation-survival rate (share of answers unchanged when the chain is deleted [28]); and the audit's own coverage (share of trace claims the tools could check). The last field is the honest one—it prices the report's confidence—and its absence in current practice is part of the blind spot, since an unaudited 100% EGR and a fully audited one are indistinguishable in today's tables. The incentive point deserves emphasis: benchmarks do not merely measure—they select. A leaderboard that reports only P(outcome correct) trains the field's collective attention on the SC ∪ FC mixture; one that reports CER alongside it makes compensated reasoning a visible, comparable defect. Evaluation is the cheapest leverage point, and the only one whose audience includes every other point on this map.

6. Discussion

6.1. The Tractability Asymmetry

The survey's recurring asymmetry deserves its own statement: multimodal process errors are externally checkable in ways text reasoning errors are not. A hallucinated object has a detector; a misread chart value has an OCR-and-parsing pipeline; a fabricated spatial relation has a grounding model. Text-side process supervision must lean on learned PRMs whose own reliability is the open question [36,48]; the multimodal setting can lean on perception tooling whose failure modes are comparatively characterized. This is why constructions (i) and (ii) of Section 5.3 use no science beyond assembly. The asymmetry also suggests where multimodal process supervision will not be hard where text was: annotation cost, the historical bottleneck of process supervision [48], can be partly replaced by tool verdicts, which are cheap, consistent, and auditable in turn. The cost comparison is roughly: a human process label costs minutes of expert time per step; a tool verdict costs milliseconds and produces a diffable record; a model-based judge costs an inference call and inherits the judge's own blind spots. Tooling is weakest exactly where perception is hardest—the primitive failures of Table 2 resist cheap verifiers—so the practical program is a coverage ladder: tool-audit the structured claims (most of the FC mass on document and chart tasks), model-judge the rest, reserve human review for calibration of the other two. If the multimodal reasoning wave adopts this ladder early, faithful multimodal reasoning could become the setting where process supervision is cheaper than outcome supervision plus its failures—not a luxury good.

6.2. The Monitoring Tension

A second-order effect deserves caution. If outcome-only RL makes unfaithful chains (Section 4.3's obscuration result [37]), then process monitoring introduced without process reward may select for chains that evade monitoring—the text-side result in code RL, transposed. The design implication is sequencing: process signals should be built into the reward at the same stage that process inspection becomes load-bearing, not bolted on afterward as an expectation. The POC framework supplies the vocabulary for what such signals must distinguish; Section 5.3's constructions show they can be assembled from existing parts.

6.3. Limits of the Framework

Three caveats bound the framework's claims. First, "soundness" is stricter than human reasoning: experts skip steps, cite evidence economically, and reach correct answers by pattern recognition that no audit would reconstruct; a CER target of zero would penalize legitimate

expertise-shaped reasoning and select verbose chains. The framework's use is as a reported quantity and a reward component, not a pass/fail bar. Second, process metrics are gameable too: a model rewarded for citing checkable evidence can learn to cite verbatim-plausible evidence, and tool verdicts inherit their tools' errors; Section 4.3's overoptimization logic applies to PRMs as to ORMs [36,38]. Third, the SC/FC boundary is epistemic, not ontological—the same trace can be sound given one reading of a blurry axis and compensated given another—so CER estimates carry the audit's own uncertainty and should be reported with it. Fourth, soundness as defined here bundles three properties—faithfulness (the chain caused the answer), grounding (cited evidence matches the image), and validity (steps follow)—that come apart in principle: a trace can be perfectly grounded yet invalid, or valid yet unfaithful. The framework's quantities target the bundle differently—EGR measures grounding, probes measure faithfulness, step-checking measures validity—and conflating them in a single "process score" would hide exactly which property a training regime improved. None of these caveats weakens the core claim; each relocates it from "process is measurable" to "process is measurable well enough to be worth measuring," which the Section 2 evidence says is currently not happening at all.

7. Conclusion

This survey joined three literatures that have been solving adjacent halves of one problem. The perception-evaluation literature documented that multimodal reasoning chains routinely misread structured content, fail primitives, and fabricate evidence while remaining fluent [3,4,7,17]. The faithfulness literature supplied the measurement tools showing that chains are often not the cause of their answers [28-30]. The reward-modeling literature showed that optimization against any proxy drifts off the true objective and that outcome-only signals cannot even represent the difference [36-38]. The POC framework states the join precisely: outcome-based training and evaluation occupy one axis of a two-dimensional space, and the cell they cannot see—compensated error—is where the documented perception failures go to be rewarded.

The prescriptive summary is short because the parts exist. Make the process checkable: at inference, by extracting before computing; in training, by tool-verified rejection sampling and process-contrastive pairs that neutralize the insensitivity principle; in evaluation, by reporting compensated-error and evidence-grounding rates beside accuracy. The multimodal setting is not merely exposed to this problem but equipped for it: perception errors are the one class of reasoning errors that mundane tooling can audit, which makes faithful multimodal reasoning the most tractable instance of a problem the field has so far met mostly with cautionary tales.

What is missing is not capability but a reporting standard—the field still publishes the SC ∪ FC mixture and calls it accuracy. The compensated-error rate is one number away from every leaderboard that wants it.

References

[1] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems 35.

[2] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. In Advances in Neural Information Processing Systems 35.

[3] Wang, Z., Xia, M., He, L., Chen, H., Liu, Y., et al. (2024). CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. arXiv preprint arXiv:2406.18521.

[4] Rahmanzadehgervi, P., Bolton, L., Taesiri, M. R., & Nguyen, A. T. (2024). Vision language models are blind: Failing to translate detailed visual features into words. arXiv preprint arXiv:2407.06581.

[5] Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. arXiv preprint arXiv:2401.06209.

[6] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[7] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[8] Cobbe, K., Klimov, O., Hesse, C., Kim, D., & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

[9] Zelikman, E., Wu, Y., Mu, J., & Goodman, N. D. (2022). STaR: Bootstrapping reasoning with reasoning. In Advances in Neural Information Processing Systems 35.

[10] Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., et al. (2024). Tulu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124.

[11] Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., et al. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948.

[12] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. In ACM Transactions on Information Systems.

[13] Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., & Fei, H. (2025). Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605.

[14] Masry, A., Long, D. X., Tan, J. Q., Joty, S., & Hoque, E. (2022). ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Findings of the Association for Computational Linguistics: ACL 2022.

[15] Methani, N., Ganguly, P., Khapra, M. M., & Kumar, P. (2020). PlotQA: Reasoning over scientific plots. In Proceedings of the AAAI Conference on Artificial Intelligence.

[16] Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., & Gao, J. (2024). MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In Advances in Neural Information Processing Systems 36

Datasets and Benchmarks Track.
[17] Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. (2024). HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[18] Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Zhang, J., Yan, M., Zhang, J., & Sang, J. (2024). AMBER: An LLM-free multi-dimensional benchmark for MLLM hallucination evaluation. arXiv preprint arXiv:2311.07397.
[19] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. In Nature Machine Intelligence, 2, 665-673.
[20] McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).
[21] Jia, R., & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[22] Zhang, J., Khayatkhoei, M., Chhikara, P., & Ilievski, F. (2025). MLLMs know where to look: Training-free perception of small visual details with multimodal LLMs. arXiv preprint arXiv:2502.17422.
[23] Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[24] Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., & Yao, H. (2023). Analyzing and mitigating object hallucination in large vision-language models. arXiv preprint arXiv:2310.00754.
[25] Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., et al. (2024). Woodpecker: Hallucination correction for multimodal large language models. arXiv preprint arXiv:2310.16045.
[26] Favero, A., Zare, Z., Zarei, A., & Morency, L.-P. (2024). Multi-modal hallucination control by visual information grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
[27] Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., et al. (2023). Faith and fate: Limits of transformers on compositionality. In Advances in Neural Information Processing Systems 36.
[28] Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., et al. (2023). Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702.
[29] 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. In Advances in Neural Information Processing Systems 36.
[30] Pfau, J., Merrill, W., & Bowman, S. R. (2024). Let's think dot by dot: Simple bounds on hidden computation in transformer language models. arXiv preprint arXiv:2404.15758.
[31] Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Faithful chain-of-thought reasoning. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL).
[32] Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Zhao, M., & Sun, M. (2024). RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR).
[33] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In International Conference on Learning Representations (ICLR).
[34] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems 35.
[35] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., et al. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.
[36] Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. In Proceedings of the 40th International Conference on Machine Learning (ICML).
[37] Baker, B., Huizinga, J., Gao, L., Dou, Z., Guan, M., et al. (2025). Monitoring reasoning models for misbehavior and the risks of promoting obfuscation. arXiv preprint arXiv:2503.11926.
[38] Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. In Advances in Neural Information Processing Systems 35.
[39] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
[40] Stroebl, B., Kapoor, S., & Narayanan, A. (2024). The limits of inference scaling through resampling. arXiv preprint arXiv:2411.17501.
[41] Yu, T., Zhang, H., Li, Q., Xu, Q., Yao, Y., et al. (2024). RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. arXiv preprint arXiv:2405.17220.
[42] Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., et al. (2024). LLaVA-CoT: Let vision language models reason step-by-step. arXiv preprint arXiv:2411.10440.
[43] Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., et al. (2025). Vision-R1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749.
[44] Yu, E., Lin, K., Zhao, L., et al. (2025). Perception-R1: Pioneering perception policy with reinforcement learning. arXiv preprint arXiv:2504.07954.
[45] Wang, W., Gao, Z., Chen, L., et al. (2025). VisualPRM: An effective process reward model for multimodal reasoning. arXiv preprint arXiv:2503.10291.
[46] Zhang, L., Hu, A., Xu, H., Yan, M., Xu, Y., Jin, Q., Zhang, J., & Huang, F. (2024). TinyChart: Efficient chart understanding with visual token merging and program-of-thoughts learning. arXiv preprint arXiv:2404.16635.
[47] Gupta, T., & Kembhavi, A. (2023). Visual programming: Compositional visual reasoning without training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
[48] Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., & Cobbe, K. (2024). Let's verify step by step. In International Conference on Learning Representations (ICLR).
[49] Uesato, J., Kushman, N., Kumar, R., Song, F., Siegel, N., Wang, L., Creswell, A., Irving, G., & Higgins, I. (2022). Solving math word problems with process- and outcome-based feedback. arXiv preprint arXiv:2211.14275.
[50] Xia, R., Zhang, B., Ye, H., et al. (2024). ChartX & ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. arXiv preprint arXiv:2402.12185.
[51] Chen, J., Kong, L., Wei, H., Liu, C., & Ge, Z. (2024). OneChart: Purify the chart structural extraction via one auxiliary token. arXiv preprint arXiv:2404.09987.
[52] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023).

Direct preference optimization: Your language model is secretly a reward model. In Advances in Neural Information Processing Systems 36.