# EVIDENCE LEDGERS: AUDITABLE CLAIM-LEVEL GROUNDING FOR VISION–LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Vision–language model answers often contain several factual claims, but standard generation provides no reliable record of the input evidence that supports each claim. We introduce *evidence ledgers*, which align evidence atoms, typed claims, and answer units through two routes, a direct evidence-to-answer alignment and one composed through the claims, with a null source that absorbs claim mass not assigned to the available evidence. Training combines a claim-level unlikelihood term weighted by the null allocation, a counterfactual swap loss, and consistency between the two routes on the model's own drafts; inference emits a claim-level *certificate* whose SUPPORTED verdict is calibrated by learn-then-test to control the false-support rate among SUPPORTED claims, provided that calibration and deployment claims follow the same distribution. On Qwen3-VL-235B the strict, human-audited unsupported-claim rate falls from $11.0\%$ to $4.6\%$. The corresponding rates are $7.2\%$ for on-policy grounder-DPO and $7.0\%$ for GRPO trained with the same grounding signal at matched compute, and $6.8\%$ for grounder-DPO with tool outputs in context; HallusionBench accuracy rises by $10.4$ points, and the gains persist on long-form hallucination benchmarks and relabelled POPE and replicate on InternVL3.5-38B. Controlled comparisons assess the contributions of the loss terms: the unlikelihood term contributes most, the swap and alignment terms add smaller increments, and coupling claims through shared evidence capacity adds under one point overall. In a study using identical answers, annotators reading ledger certificates needed $19.7\%$ less verification time than those reading grounder certificates.

## 1 INTRODUCTION

Ask a vision–language model (VLM) what a chart shows and it answers fluently: *"revenue declined after 2021, with the sharpest drop in Q3."* The sentence commits to a chart region, an axis convention, a numerical trend, and a temporal comparison; a model that is $90\%$ accurate at the answer level can be wrong about any one of these while sounding right about all. Current VLMs (Zhu et al., 2024; Peng et al., 2024; You et al., 2024; Bai et al., 2025; Wang et al., 2025), built from transformers, sparse experts, and low-rank adapters (Dosovitskiy et al., 2021; Shazeer et al., 2017; Hu et al., 2022), produce such commitments token by token and do not routinely provide an explicit mapping from each generated claim to supporting input evidence. Attribution methods (Selvaraju et al., 2017; Sundararajan et al., 2017; Lundberg and Lee, 2017) surface activations without naming a proposition, attention is an unreliable explanation (Jain and Wallace, 2019; Darcet et al., 2024), and rationales share the answer's generative prior (Jacovi and Goldberg, 2020; Wang et al., 2023b). Decoding-time methods (Chuang et al., 2024; Leng et al., 2024; Huang et al., 2024) intervene during generation and post-hoc methods (Zhou et al., 2024b; Yin et al., 2024; Asai et al., 2024) revise or assess completed drafts; grounding detectors (Kamath et al., 2021; Liu et al., 2023) localise named phrases. None of these produces, for each generated claim, an assessment tied to the evidence that supports it.

**Evidence ledgers.** An EVIDENCE LEDGER links evidence atoms such as regions, OCR lines, chart elements, and frames, typed claim atoms parsed from the answer, and answer units through three sparse entropic transport plans: a *direct* evidence-to-answer plan $\Pi^{\mathrm{EA}}$ and a plan *composed* through the claim layer, $\Pi^{\mathrm{EC}} \to \Pi^{\mathrm{CA}}$. We say that an atom *funds* a claim when the plan assigns it part of the claim's budget; funding is a transport allocation, not an entailment judgement. Evidence atoms carry



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

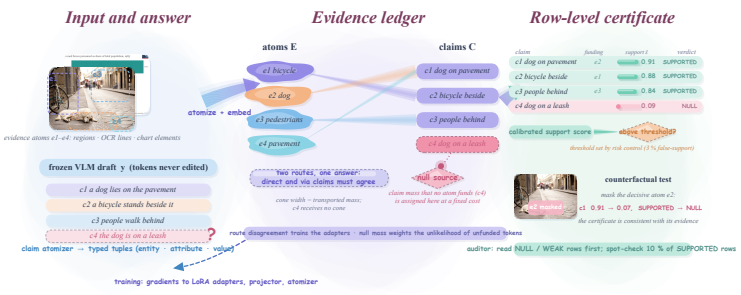



Figure 1: **A certificate for every claim.** Left: an answer-only VLM describes a street photograph fluently; one clause, "the dog is on a leash", is plausible but unverifiable. Middle: the ledger atomizes the image into evidence atoms $e_1$–$e_4$ and the answer into claims $c_1$–$c_4$; cone width is transported mass, and mass that no atom takes up is assigned to the null source. Right: the certificate is row-level. $c_1$–$c_3$ are SUPPORTED with $\hat{s} \geq 0.84$; $c_4$ receives $\hat{s} = 0.09$ and is flagged NULL while the answer text stays unchanged.

capacity upper bounds, claim and answer budgets are met exactly, and a null source with constant cost absorbs claim mass that no atom takes up more cheaply. The two routes are required to agree; their disagreement, the marginal residual, and a degree penalty train a LoRA-adapted generator, while stop-gradient null shares weight an unlikelihood term on the tokens of unfunded claims. At inference, the system solves the alignment problem once for the completed answer and emits one certificate row per claim, listing its funding atoms, shares, residual, and verdict, without editing the answer; Figure 1 shows an example.

**Findings.** First, the solver is analytically transparent: whenever no evidence capacity binds, which holds on 94% of benchmark inputs, each column of the plan reduces to a softmax over candidate atoms and the null source (Proposition 1); a control built from that construction performs within 0.5 pp of the full model, and a penalty-matched comparison credits the coupled solve with 1.3–1.5 pp only where capacities bind. Second, controlled comparisons assess the loss terms: ledger-derived and grounder-derived unlikelihood weights yield similar UCR values, removing the swap loss produces the largest increase in UCR among the loss terms, and two-route consistency adds a consistent but modest increment of 0.7–1.0 pp over its nearest controls. Third, the certificate is useful at controlled risk: at a 3% false-SUPPORTED rate it labels 1.8–2.1 pp more claims SUPPORTED than a classifier trained on the same annotations, and annotators who see it need 19.7% less verification time while finding more unsupported claims.

**Contributions.** (1) A claim-level grounding objective, null-weighted unlikelihood, counterfactual swaps, and two-route evidence alignment with a null source, whose components are assessed by seven controls trained on the same drafts, tools, and annotations; (2) a certificate whose SUPPORTED verdict is calibrated by learn-then-test risk control (Angelopoulos et al., 2021) and whose certified coverage is measured on shared answers with intervals; (3) a strict slot-level measurement protocol graded by a blinded non-author audit and corroborated by four robustness analyses and relabelled POPE (Neuhaus and Hein, 2025); and (4) a replication on a second vision stack. The main text specifies the learning objective (Section 3), the controlled comparisons used to assess component contributions (Section 5.1), and the statistical estimands (Section 4); the appendix documents the proofs (Appendix E), the experimental settings (Appendix C), and the evaluation definitions and protocols (Appendices B and D). These descriptions support independent examination of the theoretical claims and implementation of the evaluation protocol.

## 2 Related Work

**Hallucination mitigation and fine-grained feedback.** POPE (Li et al., 2023) and Hallusion-Bench (Guan et al., 2024) expose object- and reasoning-level hallucination, which contrastive decoding (Chuang et al., 2024; Leng et al., 2024), over-trust penalties (Huang et al., 2024), robust instruction tuning (Liu et al., 2024), revision pipelines (Zhou et al., 2024b; Yin et al., 2024), and preference optimisation (Zhao et al., 2023; Zhou et al., 2024a; Yu et al., 2025) reduce. Segment- and

phrase-level feedback is the closest line to our token channel: RLHF-V (Yu et al., 2024), LLaVA-RLHF (Sun et al., 2024), FDPO (Gunjal et al., 2024), HSA-DPO (Xiao et al., 2025), EFUF (Xing et al., 2024), and HALVA (Sarkar et al., 2025) learn from span- or phrase-level corrective signals; Control A of Section 5.4 is a detector-weighted instance of this family, and our GRPO baseline follows Shao et al. (2024). Our focus is on combining claim-level evidence assignments with calibrated support verdicts and on measuring their coverage at a controlled false-support rate.

**Alignment by transport and cycle consistency.** Entropic Sinkhorn (Cuturi, 2013; Peyre and Cuturi, 2019) aligns words to regions in UNITER (Chen et al., 2020b) and graphs in graph OT (Chen et al., 2020a); SuperGlue (Sarlin et al., 2020) adds a dustbin row, and partial and unbalanced transport (Figalli, 2010; Chizat et al., 2018) allow inequality marginals. Cycle consistency between direct and composed correspondences is an established signal in image translation (Zhu et al., 2017), correspondence from video (Wang et al., 2019), and temporal alignment (Dwibedi et al., 2019), and multi-marginal transport composes couplings (Pass, 2015); prompt learning with transport aligns visual and textual prompts (Chen et al., 2023); ledgers bring the cycle signal to claim-level grounding, where a direct and a composed evidence-to-answer route are required to agree.

**Claim-level factuality and risk control.** FActScore (Min et al., 2023) and self-critique (Asai et al., 2024) audit text at the claim level, attribution evaluation (Rashkin et al., 2023) and citation generation (Gao et al., 2023) attach sources to text, fine-grained reward modelling (Yan et al., 2024) scores sub-sentence correctness in VLMs, CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2023a), MMHal-Bench (Sun et al., 2024), and FaithScore (Jing et al., 2023) score generated descriptions, and RePOPE (Neuhaus and Hein, 2025) corrects POPE's labels. Our guarantee is a selective-prediction risk-control statement (Geifman and El-Yaniv, 2017; Angelopoulos et al., 2021; 2024; Guo et al., 2017), the multimodal counterpart of conformal factuality for text (Mohri and Hashimoto, 2024; Quach et al., 2024; Cherian et al., 2024).

# 3 METHOD

## 3.1 ATOMS, CAPACITIES, AND BUDGETS

Given an input $x$, evidence tools and a frozen encoder produce evidence atoms $E = \{e_i\}_{i=1}^{n}$, each with a modality tag, a location, and an embedding $z_i^E = h_\psi^E(e_i) \in \mathbb{R}^d$ from a trainable projector. The decoder generates an answer, which is divided into sentence-level units $A = \{a_k\}_{k=1}^{m}$, the sentences that contain at least one parsed claim, and embedded as $z_k^A$ by a linear map of mean-pooled decoder states, so that gradients reaching $z^A$ reach the adapters. A trainable text-side claim atomizer $h_\psi^C$, a span tagger over the decoded answer text, parses the answer into typed claims $C = \{c_j\}_{j=1}^{\ell}$ (entity, attribute, value) with embeddings $z_j^C \in \mathbb{R}^d$ (Appendix C). The demand side is normalised: answer budgets $q \in \Delta^m$, $q_k \propto \mathrm{len}(a_k)$, and claim budgets $b \in \Delta^\ell$, $b_j = \sum_{k \in \mathrm{span}(c_j)} q_k / n_k$, where $n_k$ is the number of claims whose span touches unit $k$, so that the claims of a unit share its budget equally and $\sum_j b_j = \sum_k q_k = 1$; budgets are allocation weights, not probabilities of correctness. The supply side is not: each atom has a capacity $u_i = \min(1, \mathrm{conf}_i \, w_{\mathrm{mod}}(i)) \in (0, 1]$, where $\mathrm{conf}_i$ is the producing tool's confidence and $w_{\mathrm{mod}} \in [0.8, 1.2]$ a per-modality weight, and total capacity exceeds the unit demand by a median factor of 17.9. A null source $\varnothing$ of free capacity, a dustbin in the sense of SuperGlue (Sarlin et al., 2020), is priced by a constant cost $\nu_0 = 0.62$, fit once by maximising the raw funding-set $F_1$ against adjudicated evidence on the scorer-fit claims, computed before any verdict is assigned so that no threshold enters, and then frozen.

## 3.2 THREE PLANS, TWO ROUTES, AND THE GAP

The ledger (Figure 2) consists of three non-negative sparse plans and a composed path,

$$\Pi^{\mathrm{EC}} \in \mathbb{R}_+^{(n+1)\times \ell}, \quad \Pi^{\mathrm{CA}} \in \mathbb{R}_+^{\ell \times m}, \quad \Pi^{\mathrm{EA}} \in \mathbb{R}_+^{(n+1)\times m}, \quad \widehat{\Pi}^{\mathrm{EA}} = \Pi^{\mathrm{EC}} \, \mathrm{diag}(b+\epsilon_b)^{-1} \Pi^{\mathrm{CA}}, \tag{1}$$

where the extra row is the null source and $\epsilon_b = 10^{-6}$ is a numerical guard. Feasibility is that of partial transport (Figalli, 2010; Chizat et al., 2018): $(\Pi^{\mathrm{EC}})^\top \mathbf{1} = b$, $\Pi^{\mathrm{CA}} \mathbf{1} = b$, $(\Pi^{\mathrm{CA}})^\top \mathbf{1} = q$, $(\Pi^{\mathrm{EA}})^\top \mathbf{1} = q$; concrete rows obey $(\Pi^{\mathrm{EC}} \mathbf{1})_i \le u_i$ and $(\Pi^{\mathrm{EA}} \mathbf{1})_i \le u_i$; the null row is unconstrained. With $\epsilon_b = 0$ and exact input marginals the composed plan has the target marginals; the implemented

plan is approximately marginal-consistent, and $\Gamma$ below penalises the residual. The two routes are required to agree, $\Pi^{\mathrm{EA}} \approx \widehat{\Pi}^{\mathrm{EA}}$. The training signal is the gradient-carrying *ledger gap* together with a gradient-free null term:

$$\Delta^{\nabla}_{\mathrm{led}} = \underbrace{\|\Pi^{\mathrm{EA}} - \widehat{\Pi}^{\mathrm{EA}}\|_1}_{\Delta_{\mathrm{cyc}}} + \underbrace{\lambda_g \Gamma}_{\Delta_{\mathrm{marg}}} + \underbrace{\lambda_s \Omega_s}_{\Delta_{\mathrm{sparse}}}, \qquad \Delta_{\mathrm{null}} = \lambda_{\varnothing}\, \mathrm{sg}\Big(\sum_j \Pi^{\mathrm{EC}}_{\varnothing j} + \sum_k \Pi^{\mathrm{EA}}_{\varnothing k}\Big), \quad (2)$$

so that $\Delta_{\mathrm{led}} = \Delta^{\nabla}_{\mathrm{led}} + \Delta_{\mathrm{null}}$. At convergence $\Delta^{\nabla}_{\mathrm{led}} = 0.027$ and $\Delta_{\mathrm{null}} = 0.081$ (components in Appendix B); every "Gap" in a table denotes $\Delta^{\nabla}_{\mathrm{led}}$, and $\Delta_{\mathrm{cyc}} + \Delta_{\mathrm{null}}$ is the label-free diagnostic of Section 5.5. $\Gamma$ is the $\ell_1$ violation of the four equality constraints plus the positive part of the two capacity overflows, which is $0.006$ at convergence. $\Omega_s$ is a differentiable degree-overflow penalty on soft row and column degrees with caps $K_r = 2$ and $K_c = 8$; Appendix E, Eq. (7), defines it.

### 3.3 Costs, the solver, and its decoupled form

Costs combine semantic mismatch, modality reliability, and a geometric prior, $D^{EC}_{ij} = 1 - \cos(z^E_i, z^C_j) + \rho_{\mathrm{mod}}(i) + \rho_{\mathrm{geo}}(i, j)$, where $\rho_{\mathrm{mod}}$ penalises low tool confidence and $\rho_{\mathrm{geo}}$ the distance to the claim's grounded entity (Appendix C); $D^{CA}$ and $D^{EA}$ are defined analogously. Each claim column has a candidate set $S_j$ of its $K = 8$ nearest atoms plus up to $K_{\mathrm{prior}} = 4$ protected prior edges whose label lexically matches the claim's entity ($|S_j| \le K_{\mathrm{eff}} = 12$; Appendix C). On this support we solve entropic OT (Cuturi, 2013; Peyre and Cuturi, 2019) with $\varepsilon = 0.05$ by $T = 15$ Sinkhorn iterations: semi-relaxed for $\Pi^{\mathrm{EC}}$ and $\Pi^{\mathrm{EA}}$ (Gibbs kernel $G = e^{-D/\varepsilon}$; exact column scaling $v \leftarrow b/(G^\top a)$, clipped row scaling $a_i \leftarrow \min(1, u_i/(Gv)_i)$, unscaled null row) and balanced for $\Pi^{\mathrm{CA}}$, whose row and column marginals are $b$ and $q$; its support always contains each claim's own span units, and with the budget rule of Section 3.1 the plan $\Pi^{\mathrm{CA}}_{jk} = q_k/n_k$ on those edges has exactly these marginals, so the problem is feasible on every support that contains them. Both iterations converge to their unique optima (Lemma 1; the $T = 15$ plans are within $1.2 \times 10^{-3}$ of them in $\ell_1$), selection indices are stop-gradient, and the iterations are unrolled.

**Proposition 1** (Decoupling under slack capacities)**.** *If no concrete row binds at the fixed point ($a_i = 1$ for all $i \le n$), then for every claim $j$ the column of $\Pi^{\mathrm{EC}}$ is $\Pi^{\mathrm{EC}}_{ij}/b_j = \mathrm{softmax}_{i \in S_j \cup \{\varnothing\}}(-D_{ij}/\varepsilon)$ with $D_{\varnothing j} = \nu_0$, independently of every other claim; the same holds for $\Pi^{\mathrm{EA}}$. Coupling across claims enters only through binding rows.*

The proof follows from the scaling form (Appendix E); Control B1 of Section 5.4 replaces the solver by this softmax, so its plans coincide with the ledger's wherever capacities are slack. At convergence $2.3\%$ of concrete rows have $a_i < 1$, $6.1\%$ of examples contain at least one such row, and the mean $\ell_1$ distance between a solved column and its softmax is $0.004$ overall and $0.31$ on binding columns. Median sizes are $n = 23$, $\ell = 3$, $m = 3$; the solver adds latency equal to $4\%$ of the backbone's inference time within a total overhead of $18\%$ (Proposition 4).

### 3.4 Learning objective

The objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{gold}} + \mathcal{L}_{\mathrm{ul}} + \alpha\, \Delta^{\nabla}_{\mathrm{led}} + \beta\, \mathcal{L}_{\mathrm{cf}} + \gamma\, \mathcal{L}_{\mathrm{ver}} + \eta\, \mathcal{L}_{\mathrm{cal}}, \qquad (\alpha, \beta, \gamma, \eta) = (1.0, 0.5, 0.5, 0.25), \quad (3)$$

and each loss term is applied to specified model components. **Token channel.** $\mathcal{L}_{\mathrm{gold}}$ is the teacher-forced likelihood of the gold answer and $\mathcal{L}_{\mathrm{ul}} = -\sum_j w_j \sum_{t \in \mathrm{tok}(c_j)} \log(1 - p_\theta(y_t \mid y_{<t}, x))$ is an unlikelihood term (Welleck et al., 2020) on the model's own greedy draft $y$, with weights $w_j = \mathrm{sg}((n^{\mathrm{alloc}}_j - \tau_{\mathrm{ul}})^+/(1 - \tau_{\mathrm{ul}}))$, $\tau_{\mathrm{ul}} = 0.30$, and $\mathrm{tok}(c_j)$ the tokens of $c_j$ that belong to no claim whose null share is at most $\tau_{\varnothing}$ under the current plans; no scorer or risk-control threshold enters training. Only these two losses are defined directly on token probabilities; the representation losses affect generation indirectly through the adapters. **Representation channel.** $\Delta^{\nabla}_{\mathrm{led}}$ and $\mathcal{L}_{\mathrm{cf}}$ reach the adapters through $z^A$ and the projector and atomizer through $z^E$ and $z^C$; $\Delta_{\mathrm{cyc}}$ pulls answer-unit states towards the atoms that fund them. **The null term carries no gradient.** Back-propagating the null mass would move $z^C_j$ of an unfunded claim towards its nearest candidate and hide unfunded mass without changing a token, so $\Delta_{\mathrm{null}}$ is a diagnostic and the source of $w_j$; Appendix Table G.11 shows the degenerate variant. $\mathcal{L}_{\mathrm{cf}}$ replaces the embedding of a claim's decisive atom by that of a

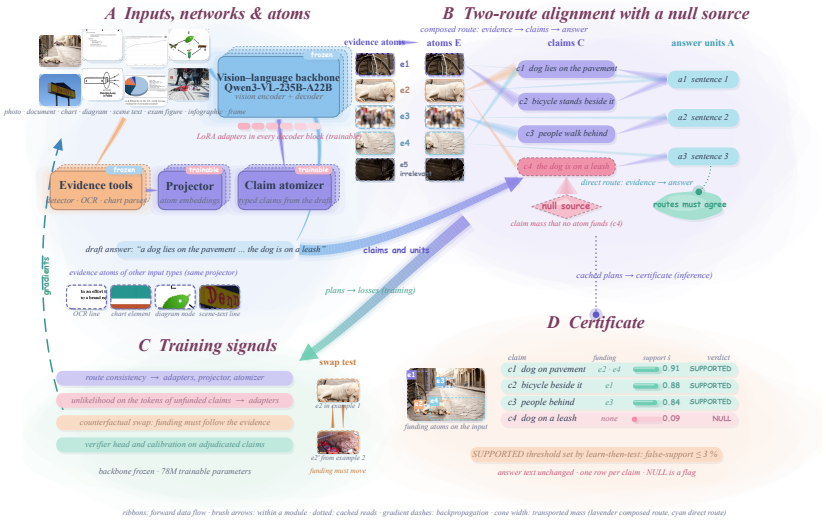



Figure 2: **Ledger architecture.** *A, inputs, networks, and atoms*: frozen evidence tools and the trainable projector $h_\psi^E$ yield evidence atoms $E$ from every input type; the frozen backbone with LoRA adapters in every decoder block drafts the answer, whose units $A$ are embedded from decoder states and whose text the atomizer $h_\psi^C$ parses into claims $C$. *B, two-route alignment*: the composed route $E \to C \to A$ and the direct route $E \to A$ are sparse entropic plans that must agree; claim mass that no atom takes up, here $c_4$, is assigned to the null source. *C, training signals*: route consistency and the swap term reach adapters, projector, and atomizer, the unlikelihood term reaches the adapters through the tokens of unfunded claims, and the verifier head is fit on adjudicated claims. *D, certificate*: one row per claim with funding atoms, support $\hat{s}_j$, and a verdict whose SUPPORTED threshold is set by learn-then-test.

same-modality atom from another example, re-solves the plans, and penalises the budget-normalised funding the swapped-in atom receives on the composed and on the direct route (Appendix C, Eq. (6)); $\mathcal{L}_{\text{ver}}$ is a cross-entropy on the verifier head $v_\omega$, which reads $(z_j^C, \{z_i^E\}_{i \in S_j}, \Pi^{\text{EC}}_{:,j})$ and outputs a raw support logit $u_j$, and $\mathcal{L}_{\text{cal}}$ is a Brier loss on $\sigma(u_j)$, both against adjudicated claims. We freeze the pretrained Qwen3-VL-235B-A22B weights (Bai et al., 2025) and train the projector and atomizer with 42M parameters, $v_\omega$ with 11M, and rank-16 LoRA adapters on the query and value projections of all 94 decoder blocks with 25M, which is 78M trainable parameters in total; Appendix C gives the details. Loss weights come from a 2B development sweep; a factor-two sweep of every weight changes UCR by at most 0.7 pp.

### 3.5 CERTIFICATE EMISSION

Emission runs once over the finished draft: the adapted decoder generates greedily, the atomizer parses claims, the solver runs, and one row per claim is emitted; nothing is edited or reordered, a NULL row is a flag on a claim that stays in the answer, and no result in the main evaluation applies a certificate-based abstention policy (Appendix M reports an optional re-decoding experiment). Shares are normalised by the claim budget and renormalised to sum to one: the concentrated share $\bar{c}_j$ over atoms with $\Pi^{\text{EC}}_{ij}/b_j \geq \zeta = 0.10$, the funding set, the diffuse share $\bar{\delta}_j$ below $\zeta$, and the null share $n_j^{\text{alloc}}$; if $S_j = \emptyset$ then $n_j^{\text{alloc}} = 1$ and $\hat{s}_j = 0$. The residual $r_j = \sum_k \Pi^{\text{CA}}_{jk} \| \Pi^{\text{EA}}_{:,k} - \widehat{\Pi}^{\text{EA}}_{:,k} \|_1 / \sum_k \Pi^{\text{CA}}_{jk}$ measures route disagreement, not unfunded content. The support score is

$$\hat{s}_j = g_{\text{iso}}\big(w^\top \varphi_j\big), \qquad \varphi_j = \Big(u_j,\ \bar{c}_j,\ r_j,\ n_j^{\text{alloc}},\ \max_{i \in S_j} \Pi^{\text{EC}}_{ij}/b_j,\ \tilde{H}_j\Big), \tag{4}$$

where $\tilde{H}_j$ is the normalised entropy of the funding profile and the logistic $w$ and isotonic $g_{\text{iso}}$ are fit post hoc on the scorer-fit claims and frozen; nothing else is learned at emission. A *contradiction flag* fires when one funding atom with share $\geq \tau_{\text{contra}} = 0.25$ binds the claim's (entity, attribute) to a value incompatible with the claim's own; it fires on 1.8% of audited rows (its effect is quantified in Appendix K). The verdict is NULL iff $n_j^{\text{alloc}} > \tau_\varnothing = 0.50$ or the flag is on; otherwise SUPPORTED iff $\hat{s}_j \geq \hat{\tau}_{\text{sup}}$ and $r_j \leq \tau_{\text{cyc}} = 0.05$; otherwise WEAK. SUPPORTED and WEAK rows are both *asserted*

and count towards claim coverage; the emitted answer retains every claim, including those stamped NULL (Algorithm 3).

## 4 ESTIMANDS AND GUARANTEE

Let $\phi(y)$ be the parsed claims of an answer $y$ and $s(c) \in \{0, \frac{1}{2}, 1\}$ the *minimum* over a claim's typed slots of the adjudicated entailment score; a claim is *unsupported* when $s < \frac{1}{2}$, that is, when any slot is scored 0; Appendix Table D.5 reports the lenient slot-average criterion. Four estimands recur. *UCR* is $\Pr(s < \frac{1}{2})$ over the parsed claims of answers as emitted; no certificate can change it. *aUCR* is the same rate over asserted claims (SUPPORTED or WEAK) and is reported with the *claim coverage*, the share of claims asserted. The *false*-SUPPORTED *rate* $\mathrm{FSR}(\tau_{\mathrm{sup}}) = \Pr(s < \frac{1}{2} \mid \hat{v} = \text{SUPPORTED})$ is the quantity a certificate reader needs, and *coverage at controlled risk*, or certified coverage, is the share of claims stamped SUPPORTED while the FSR is controlled; the two coverages are distinct quantities, $94.1\%$ and $68.2\%$ for the ledger. All four are relative to an adjudication process with $\kappa = 0.79$ between annotators; resampling annotator disagreements leaves the test FSR within $[1.9, 2.9]\%$ (Appendix D). Certificate rows are computed on the ledger's atomization and estimands on the evaluation atomizer's; the two agree on $95.1\%$ of spans, and Appendix D maps the residue.

**Theorem 1** (Selective certificate risk control)**.** *Let the $n$ calibration claims be i.i.d. draws from $\mathcal{D}_{\mathrm{claim}}$, the uniform distribution over the pooled parsed claims of the deployed system on a held-out prompt pool, where each draw samples a prompt with replacement in proportion to its number of parsed claims and keeps one of its claims uniformly, and let the deployment claims be drawn from the same $\mathcal{D}_{\mathrm{claim}}$. Fix $\vartheta, \delta \in (0, 1)$ and a grid $\tau_1 > \cdots > \tau_G$; let $n_S(\tau_g)$ and $k(\tau_g)$ be the numbers of calibration claims stamped* SUPPORTED *at $\tau_g$ and of those adjudicated unsupported, and $p_g$ the one-sided binomial $p$-value of $H_g : \mathrm{FSR}(\tau_g) > \vartheta$. With $\hat{\tau}_{\mathrm{sup}} = \tau_{\hat{g}}$ for the largest $\hat{g}$ such that $p_1, \ldots, p_{\hat{g}} \leq \delta$, and for any support score,*

$$\Pr\left(\mathrm{FSR}(\hat{\tau}_{\mathrm{sup}}) \leq \vartheta\right) \geq 1 - \delta, \tag{5}$$

*the probability being over the calibration draw. The statement is an instance of learn-then-test (Angelopoulos et al., 2021); proof in Appendix E.*

The guarantee is marginal, holds for any thresholded scorer, and concerns deployment claims drawn as the calibration claims were; $\delta$ is the failure probability of the calibration draw, so the confidence level is $1 - \delta$. If no grid point is rejected, nothing is certified and every non-NULL row is stamped WEAK; this did not occur on any calibration set. The ledger's contribution is the coverage attained at the controlled level. With $\vartheta = 0.03$, $\delta = 0.05$, and $n = 2{,}000$ the procedure selects $\hat{\tau}_{\mathrm{sup}} = 0.50$, at which 22 of the 1,296 calibration claims stamped SUPPORTED are unsupported and $p = 0.002$; on the disjoint audited test claims the FSR is $24/1{,}031 = 2.3\%$ at $68.2\%$ coverage, and by claim type it is $1.9\%$ for perception, $2.8\%$ for OCR and numeric, $2.4\%$ for reasoning, and $2.7\%$ for world-knowledge claims. The guarantee is marginal; group-conditional control would need a calibration set of the recommended size per type, as Appendix Table D.6 discusses. Under the stress regimes the rate rises to 2.8–4.4% and recalibration is required; WEAK and NULL rows lie outside the guarantee. Appendix E adds Propositions 2–4 (non-redundant gap terms, null routing, solver cost $O((TK_{\mathrm{eff}} + K_{\mathrm{eff}}^2)N)$).

## 5 EXPERIMENTS

### 5.1 SETUP AND CONTROLLED COMPARISONS

The backbone is Qwen3-VL-235B-A22B, frozen, trained with five seeds on 1.8M instruction pairs (Wei et al., 2022; Li et al., 2024b) plus 46,000 claim-level records, of which 6,200 are human-annotated claims and 39,800 constructed swap pairs. Evaluation uses fourteen benchmarks under VLMEvalKit (Li et al., 2024a): MMMU (Yue et al., 2024), MMMU-Pro, MMBench, MMStar, DocVQA (Mathew et al., 2021), TextVQA, ChartQA (Masry et al., 2022), AI2D (Kembhavi et al., 2016), OCRBench, POPE (Li et al., 2023), HallusionBench (Guan et al., 2024), VideoMME (Fu et al., 2024), EgoSchema, and RefCOCOg; three long-form hallucination benchmarks under a captioning prompt, Object HalBench (Rohrbach et al., 2018; Yu et al., 2024), AMBER (Wang et al., 2023a), and MMHal-Bench (Sun et al., 2024); and RePOPE (Neuhaus and Hein, 2025). Every system's

Table 1: **Reliability under three graders, and cost.** Strict UCR is given under the calibrated verifier, the frozen evaluator J-ind, and the blinded human audit; aUCR is measured over asserted claims at the listed claim coverage; Cert.$F_1$ is the funding-set $F_1$. Trained systems report seed means, with hUCR on the seed-17 outputs; frozen systems use one greedy run; remaining baselines and controls are in Appendix Tables G.13–G.15 and G.11. $^\dagger$: our frozen grounder attached to that system's answers.

| System | $\text{UCR}_{\text{ver}}\downarrow$ | $\text{UCR}_{\text{J-ind}}\downarrow$ | hUCR↓ | aUCR↓ | Hall.Acc↑ | POPE $F_1\uparrow$ | Cert.$F_1\uparrow$ | Cov.↑ | Lat. |
|---|---|---|---|---|---|---|---|---|---|
| Qwen3-VL-235B-A22B (backbone) | 11.4 | 11.8 | 11.0 | 8.2$^\dagger$ | 64.2 | 93.9 | 70.6$^\dagger$ | 84.2$^\dagger$ | 1.00× |
| + on-policy grounder-DPO (own drafts, matched compute) | 7.6 | 7.8 | 7.2 | 5.6$^\dagger$ | 70.4 | 96.3 | 75.6$^\dagger$ | 88.1$^\dagger$ | 1.00× |
| + on-policy GRPO (grounder claim reward, matched compute) | 7.4 | 7.6 | 7.0 | 5.3$^\dagger$ | 70.9 | 96.4 | 75.9$^\dagger$ | 88.4$^\dagger$ | 1.00× |
| + tool outputs in context + on-policy grounder-DPO | 7.1 | 7.3 | 6.8 | 5.2$^\dagger$ | 71.2 | 96.5 | 75.8$^\dagger$ | 88.3$^\dagger$ | 1.12× |
| Answer-only LoRA + post-hoc grounder @ Cov = 94.1 (same 6.2k) | 10.8 | 11.0 | 10.3 | 8.6 | 65.0 | 94.1 | 76.1 | 94.1 | 1.27× |
| Backbone + claim classifier @ 94.1 (same 6.2k) | 11.4 | 11.8 | 11.0 | 8.3 | 64.2 | 93.9 | 76.4 | 94.1 | 1.06× |
| LEDGER-VLM (ours) | **4.9** | **5.3** | **4.6** | **2.5** | **74.6** | **97.4** | **84.2** | **94.1** | 1.18× |

answers are parsed by one evaluation atomizer, a text-only Qwen2.5-7B decomposer trained before this work and disjoint from the ledger's atomizer in data, architecture, and initialisation, and graded by a frozen Qwen2.5-VL-7B verifier, by the frozen evaluator J-ind, and by human judges. Cert.$F_1$ is the funding-set $F_1$ against adjudicated evidence; for systems without a certificate it is supplied by our frozen grounder and marked $^\dagger$. Hall.Acc is HallusionBench qAcc. Answer-only LoRA on our data mixture separates the ledger losses from fine-tuning; decoding-time mitigation with DoLa (Chuang et al., 2024) and OPERA (Huang et al., 2024) and released-recipe fine-tuning with HA-DPO, POVID, and RLAIF-V place the result against prior methods; *on-policy grounder-DPO* and *on-policy GRPO* with a claim-level reward from the same grounder, both at matched compute of 695 active H100-hours per seed and equal tuning budgets, test whether preference or reinforcement optimisation of the same signal suffices; *tool outputs in context*, alone and with grounder-DPO, tests whether tool access rather than the ledger explains the gain; the post-hoc grounder and claim classifier, trained on the same 46k records and matched to the ledger's claim coverage, separate certification from generation. Every same-backbone system receives the same prompt and visual tokens; evidence atoms never reach the ledger's decoder, and inference costs $1.18\times$ with a certificate; compute totals $\approx$ 106k active H100-hours (Appendix C.1).

### 5.2 Main results

Table 1 and Appendix Table G.10 show the pattern the mechanism predicts: saturated document and OCR tasks barely move, DocVQA by 0.2; reasoning, chart, and grounding tasks gain 1.1 to 1.6 points; $\text{Avg}_{14}$ rises from 84.97 to 86.94; and HallusionBench, up 10.4, and POPE, up 3.5, account for half of the summed per-benchmark gains. The dominant effect is on unsupported content: strict UCR falls from 11.4% to 4.9%, HallusionBench qAcc rises 3.4 points over the strongest baseline, whereas a yes/no logit-bias correction of the backbone reaches only 67.0, and POPE $F_1$ reaches 97.4. Because POPE's labels contain errors that a system can agree with, RePOPE, which corrects them, is our primary object-hallucination evidence: the ledger scores 94.7 $F_1$ against 90.3 for the backbone and 92.6–92.9 for the on-policy baselines, and it loses least under relabelling (Appendix Table G.19). On the long-form benchmarks for which the fine-tuning baselines were developed, the ledger leads every baseline; Appendix Table G.19 gives the values. Decoding-time mitigation cuts UCR by 0.8–1.5 pp at 1.1–1.9× latency, tool outputs in context reach 9.6% alone and 7.1% with grounder-DPO, and the controls trained on the same annotations leave the emitted answer at 10.8–11.4%.

### 5.3 Independent measurement: the blinded human audit

UCR was audited by human judges independently of every learned component. The same 520 prompts, stratified, prompt-paired, and disjoint from training and calibration, were used for every audited system; outputs were frozen, and non-author annotators from a disjoint pool saw randomised packets with system names, certificates, scores, and verdicts hidden. Two annotators scored each claim's typed slots as $1/0.5/0$ with $\kappa = 0.79$ and an adjudicator resolved disagreements; unjudgeable items and atomization failures, 3.1% of claims, were excluded per claim. The audit covers the seed-17 outputs of the six systems in Table 2 and the thirteen attribution generators in Table 3 on all 520 prompts; twelve further generators were audited on a 200-prompt subsample, where their hUCR values lie within 1.9 pp of the verifier's.

Table 2: **Blinded human audit** on 520 shared prompts, seed-17 outputs, strict criterion. hUCR scores every parsed claim as emitted and is claim-micro; the answer-macro values are $10.6/7.0/9.9/4.3\%$. ahUCR scores asserted claims only. Paired differences are system minus ledger; their prompt-clustered 95% intervals are given in Section 5.3.

| System | Judged claims | Unsupported | hUCR | Asserted (claim cov.) | Unsup. asserted | ahUCR | Answer cov. | Paired ΔhUCR |
|---|---|---|---|---|---|---|---|---|
| Qwen3-VL-235B | 1,498 | 165 | 11.0% | 1,498 (100%) | 165 | 11.0% | 97.5% | +6.4 pp |
| + on-policy grounder-DPO | 1,503 | 108 | 7.2% | 1,503 (100%) | 108 | 7.2% | 97.3% | +2.6 pp |
| + on-policy GRPO | 1,501 | 105 | 7.0% | 1,501 (100%) | 105 | 7.0% | 97.3% | +2.4 pp |
| + tool outputs in context + grounder-DPO | 1,504 | 102 | 6.8% | 1,504 (100%) | 102 | 6.8% | 97.3% | +2.2 pp |
| Answer-only LoRA + grounder @ 94.1 | 1,506 | 155 | 10.3% | 1,417 (94.1%) | 122 | 8.6% | 95.4% | +5.7 pp |
| LEDGER-VLM | 1,512 | **70** | **4.6%** | 1,423 (94.1%) | **35** | **2.5%** | 96.9% | 0 (reference) |

Table 3: **Attribution.** Left: generators trained on the same drafts, tools, prompt, and annotations, three seeds each; hUCR comes from the blinded audit of the seed-17 outputs, ΔhUCR is system minus ledger, and the prompt-clustered intervals are given in the text; Hall.Acc for every row appears in Appendix Table K.25. Right: three generators and three scorers on the same frozen answers, whose hUCR is 4.6, 10.3, and 11.0%, with certified coverage at $\vartheta = 3\%$, AUROC for adjudicated-unsupported claims, and two scorer variants.

| Generator | UCR | hUCR | ΔhUCR |
|---|---|---|---|
| Answer-only LoRA | 10.8 | 10.3 | +5.7 |
| $+\mathcal{L}_{\mathrm{ul}}$ only, ledger null shares | 7.0 | 6.6 | +2.0 |
| Control A: grounder-weighted $\mathcal{L}_{\mathrm{ul}}$ only | 7.2 | 6.9 | +2.3 |
| $+\Delta_{\mathrm{cyc}}$ only, via $\Pi^{\mathrm{EA}}$ | 8.8 | 8.4 | +3.8 |
| + InfoNCE alignment only | 8.3 | 8.0 | +3.4 |
| $+\mathcal{L}_{\mathrm{ul}} + \Delta_{\mathrm{cyc}}$, both channels | 6.7 | 6.3 | +1.7 |
| Control B2: softmax + dustbin, no $\Pi^{\mathrm{EA}}$ | 6.3 | 6.0 | +1.4 |
| Control B1: softmax + dustbin, $\Pi^{\mathrm{EA}}$ kept | 5.4 | 5.1 | +0.5 |
| Control C: InfoNCE in place of $\Delta_{\mathrm{cyc}}$ | 5.6 | 5.3 | +0.7 |
| LEDGER-VLM | **4.9** | **4.6** | 0 (ref.) |
| On-policy grounder-DPO | 7.6 | 7.2 | +2.6 |
| $+\Delta_{\mathrm{cyc}}$ / $+\Delta_{\mathrm{cyc}} + \mathcal{L}_{\mathrm{cf}}$ | 6.6 / 6.1 | 6.3 / 5.8 | +1.7 / +1.2 |
| $+\mathcal{L}_{\mathrm{ul}}$ / $+\Delta_{\mathrm{cyc}} + \mathcal{L}_{\mathrm{cf}} + \mathcal{L}_{\mathrm{ul}}$ | 6.3 / 5.6 | 6.0 / 5.3 | +1.4 / +0.7 |

| Answers / scorer | Ledger | Classifier | Grounder |
|---|---|---|---|
| *Certified coverage at $\vartheta = 3\%$ (%)* | | | |
| Ledger answers | **68.2** | 61.4 | 58.7 |
| Control answers | **51.9** | 49.8 | 47.6 |
| Backbone answers | **50.4** | 48.6 | 46.9 |
| *AUROC, adjudicated-unsupported claims* | | | |
| Ledger answers | **.86** | .82 | .79 |
| Control answers | **.84** | .80 | .77 |
| Backbone answers | **.83** | .80 | .77 |
| *Scorer variants, certified coverage (%)* | | | |
| $u_j$ only, ledger answers | 65.0 | – | – |
| $u_j$ only, control answers | 50.9 | – | – |
| $u_j$ only, backbone answers | 49.4 | – | – |
| Re-fit per generator, control answers | – | 50.6 | 48.9 |
| Re-fit per generator, backbone answers | – | 49.0 | 47.8 |

**Result and robustness** (Table 2; Appendix Figures A.3 and A.2). The audit reproduces the automatic numbers, $11.0\% \to 4.6\%$ against $11.4\% \to 4.9\%$, with grader agreement $\kappa = 0.83$–$0.86$; the paired differences of Table 2 carry prompt-clustered 95% half-widths of 1.1–1.2 pp; the interval is $[5.2, 7.6]$ for the backbone and $[1.1, 3.3]$ for grounder-DPO with tool outputs in context. Counting every excluded item against the ledger shrinks the ledger−backbone gap from $-6.4$ to $-3.2$ pp, with interval $[-4.5, -2.0]$; manual atomization by annotators blind to any atomizer on 250 prompts gives 11.3% for the backbone and 4.9% for the ledger; leaving any one benchmark out keeps the gap between $-5.6$ and $-7.2$ pp; and the gap holds in every benchmark family, from $-4.0$ to $-9.3$ pp.

### 5.4 ATTRIBUTION: WHICH COMPONENT CHANGES THE GENERATOR

Appendix Table G.11 gives all 35 configurations. **Token-level supervision.** The unlikelihood term is the largest single contributor: added alone to answer-only LoRA it lowers UCR from 10.8 to 7.0%, and Control A, which weights the same term by the frozen grounder instead of the ledger's null shares, reaches 7.2%; the two weightings yield similar values, with paired hUCR intervals against the ledger of $[1.0, 3.1]$ and $[1.3, 3.4]$ pp. Among the loss terms, removing the swap loss produces the largest increase in UCR, from 4.9 to 7.8%. **Representation alignment.** Alone, the cycle term lowers UCR by 2.0 pp and an InfoNCE alignment of answer-unit states to matched atoms by 2.5 pp; with the unlikelihood term they reach 6.7 and 7.1%. Control C, which replaces $\Delta_{\mathrm{cyc}}$ by InfoNCE in the full objective, costs 0.7 pp, with intervals $[0.0, 1.4]$ under the audit and $[0.2, 1.2]$ paired over seeds; removing the direct route from the softmax construction, B2 against B1, costs 0.9 pp, with intervals $[0.1, 1.8]$ and $[0.3, 1.5]$; both survive a Holm correction over the three pre-specified comparisons, as Appendix D details. **Capacity coupling.** Control B1 retains the direct route and every loss term except the row-degree component of the penalty, and replaces the coupled solver by the per-claim softmax of Proposition 1; it is within 0.5 pp of the ledger, with a prompt-clustered interval of $[-0.4, 1.5]$, as the proposition predicts. The penalty-matched comparison is the ledger without its row cap, which reaches 5.6%: relative to it, coupling gains 1.3 pp on the 6.1% of prompts with a binding row and 1.5 pp on a constructed slice of counting, duplicated-object, and shared-OCR-line prompts, and loses 0.3 pp on slack prompts (Appendix Table G.20). Two removals are worse than answer-only LoRA, no null source at 12.3% and $\Pi^{\mathrm{CA}}$ fixed to span overlap at 8.7%: without a sink for unfunded mass or a learned composition, the cycle and marginal terms align unfunded claims

with whatever atoms exist, so the transport terms help only within the full objective. The same mechanism explains why removing $\Pi^{\mathrm{EA}}$ from the full objective hurts far more than removing it from the softmax construction, $10.1$ against $6.3\%$. **Transfer to preference optimisation.** Adding $\Delta_{\mathrm{cyc}}$ to the on-policy DPO model lowers its UCR from $7.6$ to $6.6\%$, the swap loss as well to $6.1\%$, the grounder-weighted unlikelihood term alone to $6.3\%$, and all three to $5.6\%$, against the ledger's $4.9\%$. **Certificate versus generator** (Table 3, right). On shared frozen answers the ledger scorer labels more claims SUPPORTED at $\vartheta = 3\%$ than the classifier and the grounder in every row; the informative rows are the neutral ones, where the margin over the classifier is $2.1$ pp with interval $[0.7, 3.5]$ on control answers and $1.8$ pp with interval $[0.4, 3.3]$ on backbone answers; on the ledger's own answers the margin is $6.8$ pp, inflated by the generator's lower unsupported rate. Removing the explicit plan statistics from $\varphi_j$ costs $1.0$ pp of this margin, although the verifier head itself still reads the $\Pi^{\mathrm{EC}}$ column. NULL rows are weak pointers: their precision is $41\%$ against $37\%$ for the grounder at matched flags on backbone answers, and on the ledger's own answers null dominance alone has precision $26\%$, the contradiction flag supplying $19$ of $35$ hits (Appendix K). The HallusionBench gain is a reduction of the "yes" bias on "no"-gold items on slices that share no image or question with any training, annotation, or calibration set (Appendix Table K.26).

### 5.5 STRESS, SCALING, CALIBRATION, DIAGNOSTICS, AND REPLICATION

**Stress.** Under five corruption regimes at seven severities the ledger's UCR at severity $0.9$ is $13.1$–$17.7\%$ against $40.2$–$50.3\%$ for the backbone, and task success, which counts refusals as failures, is highest for the ledger in every regime (Appendix H). **Scaling.** Over a $6 \times 6$ sweep of Qwen3-VL checkpoints and token budgets UCR falls along both axes; an 8B configuration reaches $6.9\%$ UCR at $253$ H100-hours per seed. **Calibration and diagnostics.** ECE is $1.4$ against $2.3$–$2.4$ for the controls, the risk–coverage frontier lies below the grounder's at every coverage, and the label-free $\Delta_{\mathrm{cyc}} + \Delta_{\mathrm{null}}$ correlates with UCR under the frozen J-ind evaluator at Spearman $0.71$ within regimes and $0.97$ across all 35 cells (Appendix Figure A.4).

**Cross-backbone replication.** On InternVL3.5-38B (Wang et al., 2025), whose InternViT-6B vision tower, visual tokenizer, and scale differ from Qwen3-VL's, retraining all components over three seeds lowers strict UCR under J-ind from $12.6$ to $5.9$, a relative reduction of $53\%$ against $52\%$ on Qwen3, and a blinded 200-prompt manual-atomization audit gives $11.7\%$ against $5.5\%$ (Appendix Table G.12); transferring the text-side atomizer frozen and retraining only the projector and adapters reaches $6.4\%$ at $38\%$ of the compute.

### 5.6 HUMAN STUDY

Fourteen trained internal annotators who were not authors compared paired outputs on $200$ prompts per comparison, three annotators per prompt, randomised and blinded (Appendix N). In the decisive arm the answer is fixed and both certificates use one schema, so only the scorer differs: annotators preferred the ledger certificate in $64.8$, $70.2$, and $68.5\%$ of judgements for helpfulness, trust, and auditability, needed $19.7\%$ less verification time, found $81\%$ of unsupported claims against $69\%$, and identified the certificate source with $52\%$ accuracy against a chance level of $50\%$. In a seeded-error study annotators found $78\%$ of unsupported claims with certificates against $61\%$ without, but only $22\%$ of false-SUPPORTED rows, which they skip; Appendix N quantifies a spot-check remedy.

## 6 CONCLUSION

Evidence ledgers align evidence, claims, and answer units through two routes with a null source, train a generator through null-weighted unlikelihood, counterfactual swaps, and route consistency, and emit a claim-level certificate with one risk-controlled verdict. The unlikelihood term contributes most of the gain, the swap and alignment terms add smaller increments, and the coupled solver adds under one point. Under a strict, blinded audit the unsupported-claim rate more than halves, from $11.0\%$ to $4.6\%$, and on identical answers the certificate certifies more claims at a controlled $3\%$ false-support rate while cutting verification time by $19.7\%$.

## AI USE STATEMENT

During manuscript preparation, we used large language models as writing assistants to improve clarity and readability, including refining sentence structure, correcting grammar, and shortening overly verbose passages. The research ideas, methodological design, theoretical developments, experiments, and analyses were developed, implemented, and validated by the authors. All AI-assisted edits were reviewed by the authors, who retained control over the final wording and take full responsibility for the accuracy, originality, and integrity of the submission.

## ETHICS STATEMENT

Evidence Ledgers is intended to reduce unsupported multimodal claims and support human verification. The main human study involved compensated, non-author annotators under an institutional minimal-risk exemption, with written consent and rights to skip tasks and withdraw; no personal data were collected, and per-annotator logs were deleted after aggregation (Appendix N). Benchmarks and training corpora are used under their respective licenses without redistribution. Certificates can encourage over-reliance: the seeded-error study shows that annotators frequently missed unsupported claims labelled SUPPORTED. The statistical guarantee controls the false-support rate under the stated calibration assumptions; it does not guarantee the correctness of individual claims or extend to arbitrary distribution shifts. Appendices M and O discuss remaining risks and recommended deployment safeguards, including spot-checking SUPPORTED claims, human escalation for claim–evidence contradictions, monitoring, and recalibration. Consequential applications should retain human oversight.

## REPRODUCIBILITY STATEMENT

The main text specifies the learning objective and certificate construction (Section 3), statistical estimands and risk-control assumptions (Section 4), and experimental setup and controlled comparisons (Section 5.1). Appendix E provides the proofs, and Appendix F presents pseudocode for the transport solver, training procedure, and certificate emitter. Appendix C documents model checkpoints, tool versions, component architectures, data sources and splits, hyperparameters, tuning budgets, prompts, decoding settings, random seeds, and the certificate schema, together with an 8B reference configuration; Appendix C.1 details computational accounting. Appendices B and D specify evaluation metrics, statistical procedures, calibration, annotation, and audit protocols, while Appendix N details the human-study design and aggregation procedures. These descriptions support independent examination of the theoretical claims and reimplementation of the method and evaluation protocol.

## REFERENCES


Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt learning with optimal transport for vision-language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *Proceedings of the 37th International Conference on Machine Learning*, 2020a.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal image-TExt representation learning. In *European Conference on Computer Vision (ECCV)*, 2020b.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, 2024.

Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010.

Chaoyou Fu, Yuhan Dai, Yuzhe Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiang Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and Dinesh Manocha. 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*, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Annual Meeting of the Association for Computational Linguistics*, 2020.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. FaithScore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*, 2016.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Bo Li, Kaichen Wang, Renrui Zhang, Dong Xu, Hengshuang Cheng, Ziwei Liu, and Chunyuan Li. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024a.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Empirical Methods in Natural Language Processing*, 2023.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for vqa on document images. In *IEEE Winter Conference on Applications of Computer Vision*, 2021.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Yannic Neuhaus and Matthias Hein. RePOPE: Impact of annotation errors on the POPE benchmark. *arXiv preprint arXiv:2504.15707*, 2025.

Brendan Pass. Multi-marginal optimal transport: theory and applications. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1771–1790, 2015.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *International Conference on Learning Representations (ICLR)*, 2024.

Gabriel Peyre and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 2019.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations (ICLR)*, 2024.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

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

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arık, and Tomas Pfister. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *International Conference on Learning Representations (ICLR)*, 2025.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Haoran Hao, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Ying Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Hui Deng, Biqing Qi, Qipeng Guo, Wenwei Zhang, Yuzhe Gu, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Bowen Zhou, Weijie Su, Kaiming Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*, 2020.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. EFUF: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1167–1181, 2024.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. ViGoR: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision (ECCV)*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations (ICLR)*, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 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)*, 2024.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024a.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

## Appendix Contents


A Extended figures . . . 17

B Evaluation protocol and statistical accounting . . . 17

C Reproducibility manifest . . . 24

D Calibration measurements and audit protocol . . . 29

E Proofs . . . 31

F Pseudocode as executed . . . 35

G Complete tables . . . 35

H Stress regimes . . . 36

I Atomizer error analysis . . . 37

J Certificate gallery on real benchmark inputs . . . 38

K Verdict confusion, error analysis, and the task-score mechanism . . . 40

L Falsification matrix . . . 45

M Deployment policy . . . 46

N Human study protocol . . . 47

O Broader impact and corner cases . . . 47

# A Extended figures

This section gathers the composite figures referenced from the main text: stress and scaling (Figure A.1), audit robustness (Figures A.2 and A.3), calibration and diagnostics (Figure A.4), the ablation Pareto and Hinton views (Figure A.5), the cross-backbone radar (Figure A.6), the human-study radar (Figure A.7), and the real-input analyses behind the certificates (Figures A.8–A.11). All values are those of the corresponding tables; no figure introduces a number that does not appear in a table.

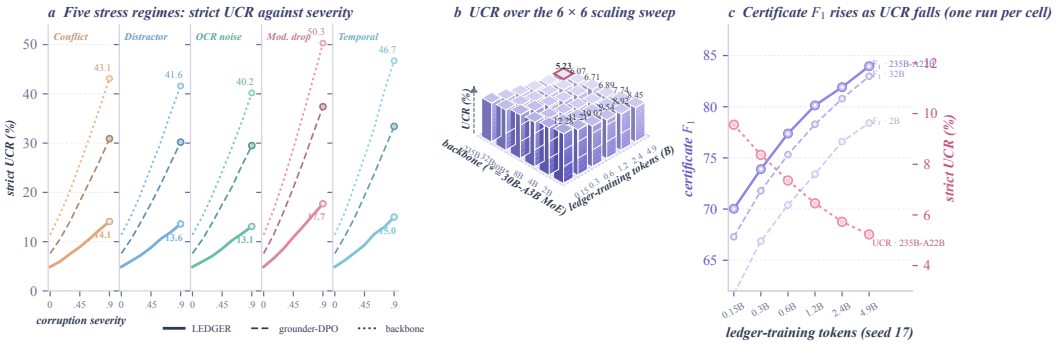



Figure A.1: **Stress and scaling.** (a) Strict UCR against corruption severity for the five regimes, one panel per regime: solid, the ledger; dashed, on-policy grounder-DPO; dotted, the backbone (values of Table H.22). (b) As-emitted UCR over the $6 \times 6$ backbone$\times$tokens sweep of released Qwen3-VL checkpoints (isometric; one run per cell; the 235B-A22B / 4.9B cell, the seed-17 main run, is outlined). (c) Dual axes: certificate $F_1$ (left; 235B-A22B solid, 32B and 2B dashed) rises as UCR (right, hollow spheres) falls with ledger-training tokens.

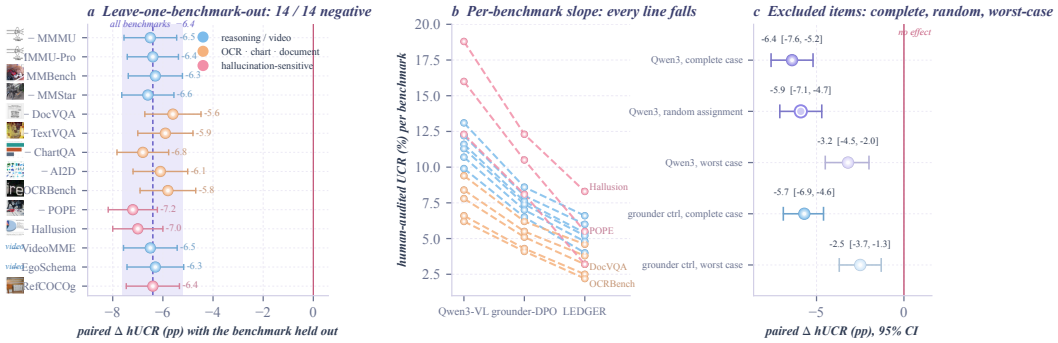



Figure A.2: **Robustness of the audit itself.** (a) Leave-one-benchmark-out forest plot of the paired Ledger−Qwen3 difference in hUCR, with a thumbnail of each held-out benchmark; the band is the all-benchmark estimate. (b) Per-benchmark slope graph from the backbone through on-policy grounder-DPO to the ledger, coloured by family: every line falls, steepest for the hallucination-sensitive family. (c) Blinded estimates versus the worst-case assignment of the 3% excluded items (all counted against the ledger, all for the comparator); both separations survive, at $-3.2$ and $-2.5$ pp, and the random-assignment estimate ($-5.9$ pp) is marked for reference.

# B Evaluation protocol and statistical accounting

**What varies across runs.** Two kinds of uncertainty are reported and never mixed. *Trained systems* (the ledger, the four controls trained on the 6.2k annotations, the noise-seeded VCD, the fine-tuning baselines LURE, HA-DPO, POVID, RLAIF-V, and the on-policy grounder-DPO and GRPO baselines) are retrained from scratch for each seed; the seed controls LoRA initialisation, data order, and dropout, and every reported value is the mean over seeds ($\{17, 29, 41, 53, 67\}$ for ours and the controls, $\{17, 29, 41\}$ for the fine-tuning baselines). For a metric with seed values $m_1, \ldots, m_S$ we report $\bar{m}$, the seed standard deviation $\mathrm{SD} = \sqrt{\frac{1}{S-1}\sum_s (m_s - \bar{m})^2}$, and $\bar{m} \pm t_{S-1;0.975}\,\mathrm{SD}/\sqrt{S}$. *Frozen systems* (API models, the unmodified backbone, DoLa, OPERA, Woodpecker) are decoded

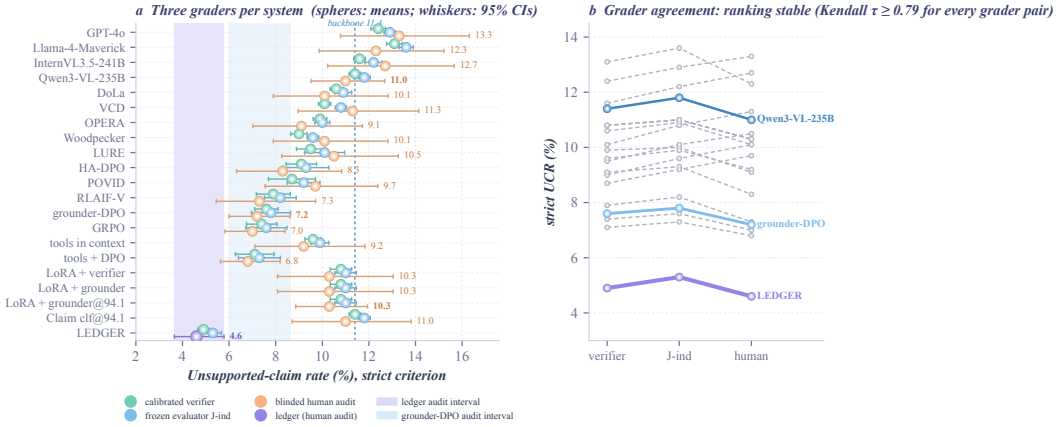



Figure A.3: **Three graders, one conclusion.** (a) UCR under the calibrated verifier (mint), J-ind (sky), and the human panel (apricot; lavender for the ledger); whiskers are seed intervals for the automatic graders and Wilson intervals for the human panel, shown for reference; the prompt-clustered intervals reported in the text are the ones used for inference. (b) Across graders neighbouring systems swap places within their intervals, but the ranking is stable overall (Kendall $\tau = 0.98$ between the two automatic graders and $\geq 0.79$ against the human panel), and the ledger's margin over the reference baseline (on-policy grounder-DPO) is 2.5–2.7 pp under all three ($\kappa = 0.83$–$0.86$).

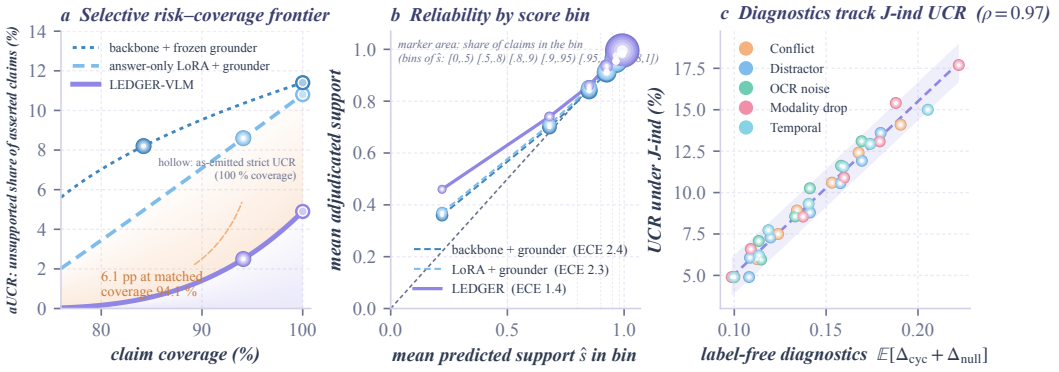



Figure A.4: **Risk–coverage, calibration, and label-free diagnostics.** (a) aUCR versus claim coverage as the assertion threshold sweeps ($\tau_{\varnothing}$ for the ledger, the abstention threshold for the grounder; $\hat{\tau}_{\text{sup}}$ and $\tau_{\text{cyc}}$ do not move this curve because they only exchange SUPPORTED and WEAK rows); spheres mark the operating points and the apricot band is the gain at matched claim coverage (6.1 pp of aUCR at 94.1%: 2.5 vs. 8.6); hollow spheres are the as-emitted strict UCR at 100% coverage. (b) Reliability of the certificate support $\hat{s}$ on the asserted rows of the 1,500-claim calibration-test set: mean adjudicated support against mean score in six score bins, with marker area proportional to the share of asserted rows in the bin; the ledger's mass concentrates in the top two bins (86% of asserted rows), and its ECE is 1.4 against 2.3–2.4 for the controls (Table D.7). (c) Each sphere is one regime×severity cell of Section 5.5; the abscissa is the mean of the online terms $\Delta_{\text{cyc}} + \Delta_{\text{null}}$ and the ordinate the UCR under the frozen J-ind evaluator; Spearman $\rho = 0.97$ over 35 cells, 0.71 within regimes after partialling out severity, and AUROC 0.79 for detecting answers with an unsupported claim on clean data with a bootstrap band. This relation is empirical and is used as an operational diagnostic, not as a bound.

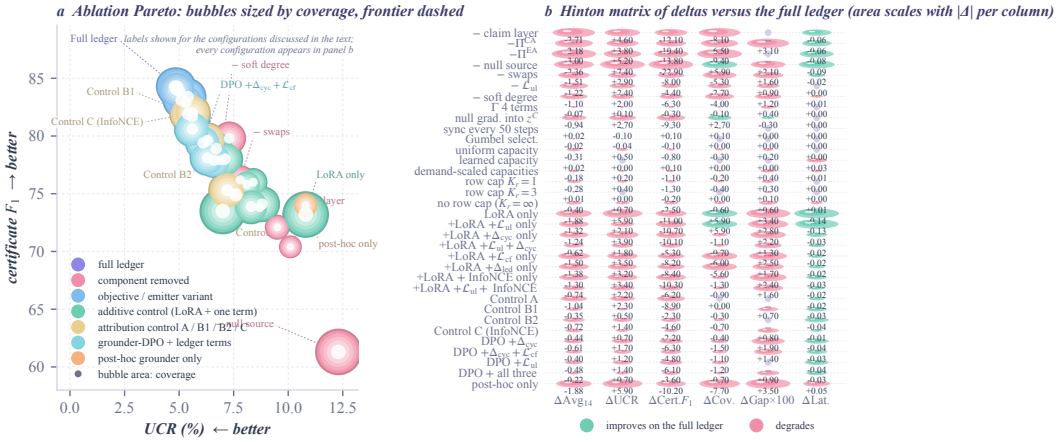



Figure A.5: **Ablation hierarchy** (values of Table G.11). (a) Bubble Pareto chart: UCR against certificate $F_1$ with bubble area proportional to coverage; coral bubbles remove a component, sky bubbles execute an objective or emitter variant, and the dashed frontier is traced by the full ledger and its variants. Removing the null source moves furthest from the frontier. (b) Hinton matrix of the deltas against the full ledger: bubble area scales with $|\Delta|$ within each column (mint improves, coral degrades) and the raw deltas are printed beneath.

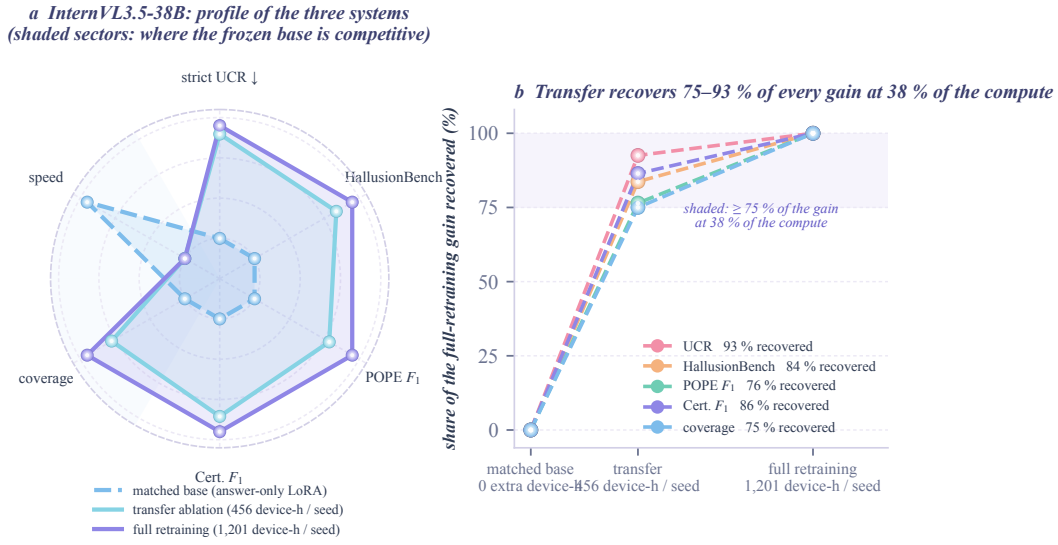



Figure A.6: **Cross-backbone replication** (values of Table G.12). (a) Radar profile of the matched answer-only base (dashed), the transfer ablation, and full ledger retraining on InternVL3.5-38B over six axes (UCR and latency inverted so that outward is better); shaded sectors mark the two axes on which the frozen base is competitive. (b) Slope graph of the share of the full-retraining gain recovered by the transfer ablation at 38% of the compute: 75–93% on every metric, so token-side modules move across backbones while hidden-state-facing modules must be retrained.

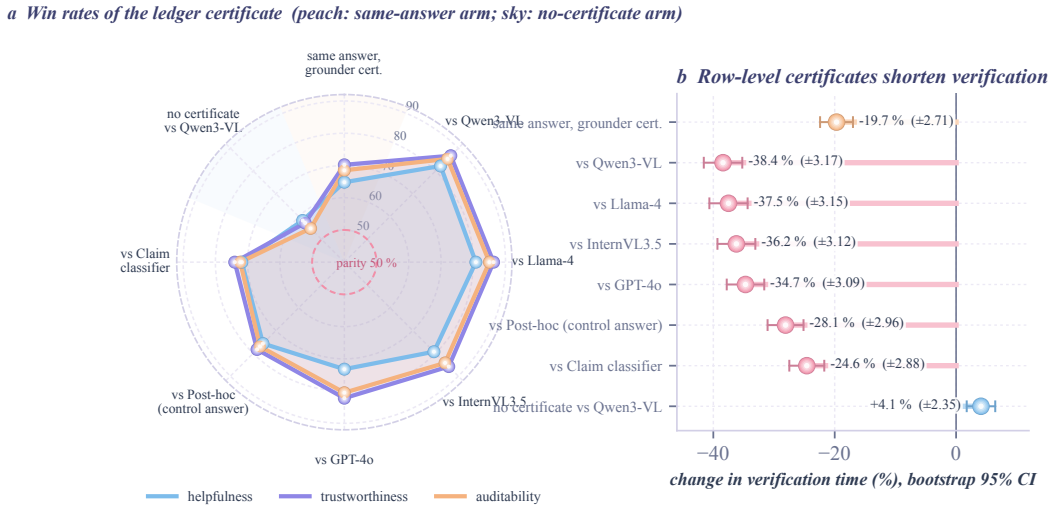



Figure A.7: **Human study.** (a) Radar chart of win rates for the ledger certificate against each comparator on helpfulness, trustworthiness, and auditability (parity ring at 50%); the same-answer arm, in which the ledger answer is shown with a ledger certificate and with a grounder certificate so that only the certificate differs, is shaded. (b) Change in verification time with bootstrap 95% whiskers; the no-certificate arm bounds the answer-only effect.

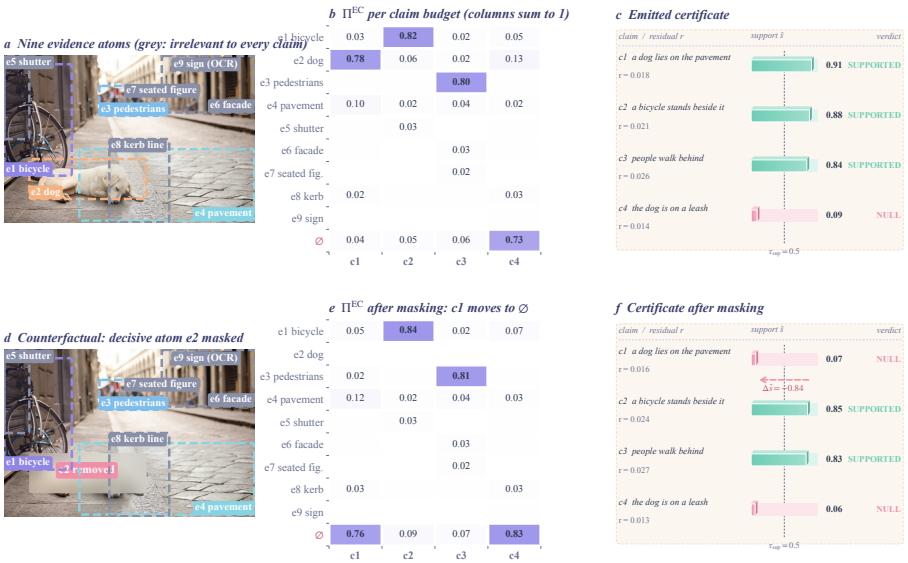



Figure A.8: **Ledger on a real photograph.** Evidence atoms on a COCO/POPE street scene fund four claims through $\Pi^{\mathrm{EC}}$; the leash claim is unfunded and its mass is assigned to the null source. Masking the decisive dog atom moves $c_1$ onto $\emptyset$ and flips its certificate from SUPPORTED to NULL ($\Delta\hat{s} = -0.84$): the certificate is counterfactually consistent with the atom it cites, which is a property of the certificate, not a causal statement about how the answer was generated.




Figure A.9: **Transport plans and certificates for three gallery inputs** (POPE, ChartQA, DocVQA). Left: evidence atoms on the real input; middle: the sparse plan $\Pi^{\mathrm{EC}}$ per claim budget with the null row $\emptyset$; right: the emitted certificate. In every case the unsupported claim (a bowl, a post-2015 trend, a signature) receives its budget from the null source and never from a distractor atom, which is the behaviour the null-source ablation of Table G.11 removes.

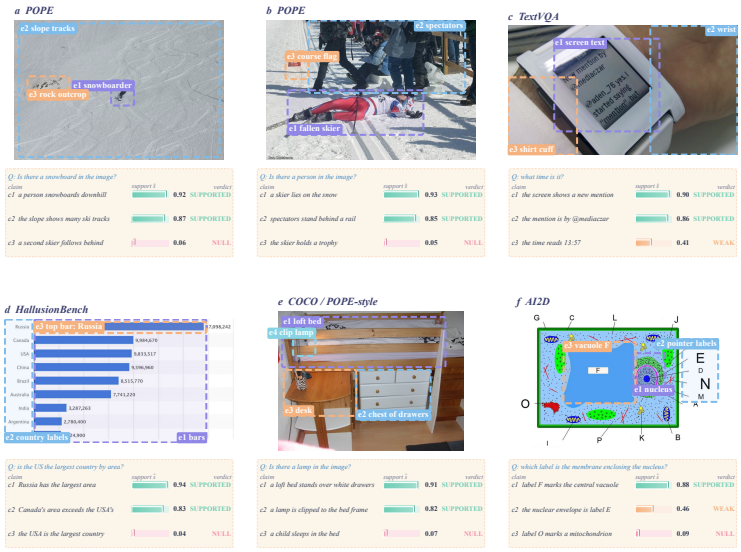



Figure A.10: **Qualitative certificates on six benchmark inputs** (POPE, POPE, TextVQA, HallusionBench, COCO/POPE-style, AI2D), each shown with its original benchmark question, the evidence atoms the ledger selected, and the emitted rows. Supported rows are funded by the boxed atoms; the WEAK rows (a partly occluded clock reading, an ambiguous diagram label) are asserted as low-confidence claims and count towards coverage; the NULL rows name the exact assertion that no atom funds (a second skier, a trophy, a child, a mitochondrion at label O), so a reader can reject one row without discarding the answer. Every value shown is read from the emitted certificate JSON, as are the gallery rows of Figure J.12.




Figure A.11: **Stress regimes on real inputs.** Top: a DocVQA memo under OCR noise and modality drop; bottom: a POPE photograph under a pasted visual distractor and a contradictory retrieved note. Under OCR noise the verdict of the affected claim moves from SUPPORTED to WEAK and its funding is never re-attached to another atom; when the image modality is removed there are no candidate atoms, so every row has $n^{\text{alloc}} = 1$, $\hat{s} = 0$, and the NULL verdict (the deployment policy of Appendix M *would* turn such a certificate into a refusal, but no number of the main evaluation applies that policy); a distractor inset becomes an atom but funds no claim about a bowl; and a retrieved note asserting a bowl becomes a text atom that *agrees* with the false claim "a bowl is present" and funds 0.31 of its budget, the modality penalty $\rho_{\text{mod}}$ keeping the share below the null-dominance threshold, so the row stays NULL by null dominance ($n^{\text{alloc}} = 0.62$) and not by the contradiction flag, which by definition cannot fire on an atom that agrees with the claim. The panel exposes a real limitation: a false text atom can fund a false claim, and contradictions *between* atoms (image versus retrieved text) are not modelled; the ledger has no notion of negative evidence, and the rise of UCR under the contradictory-retrieval regime (Appendix H) is this failure mode in aggregate. These are the row-level behaviours behind the curves of Figure A.1.

greedily at temperature 0 and produce one output per prompt; nothing varies across seeds, so they are single runs, and their intervals are prompt-level bootstrap intervals (10,000 resamples of the evaluation set) that describe sampling uncertainty over prompts. Claim-level rates on the 520-prompt audit and every paired difference carry prompt-clustered bootstrap intervals, because the claims of one answer are dependent; single-system rates on the 200-prompt grader-agreement subsample carry Wilson intervals, and the human-study win rates carry prompt-clustered bootstrap intervals over their 600 judgements. OCRBench is kept in its native score in tables and divided by ten inside $\text{Avg}_{14}$, which is the exact mean of the fourteen cells.

**Claim types and what counts as support.** Annotators label each claim as perception (objects, attributes, spatial relations), OCR/numeric (text, numbers, chart values), reasoning (a step derived from earlier claims and the input), or world knowledge (a fact not shown in the input). A claim is supported when it is entailed by the input, derivable from the input and earlier supported claims, or a standard fact that the input does not contradict; only claims that are none of these count as unsupported. Table B.1 reports UCR and the ledger's NULL rate by type. The ledger over-flags world-knowledge claims (9.2% NULL against 4.1% unsupported), which is the price of not giving the ledger a knowledge source; $\mathcal{L}_{\text{ul}}$ receives zero weight below a null share of $\tau_{\text{ul}} = 0.30$, so a flagged but true world-knowledge claim is rarely penalised in training; on the audit sample the ledger emits 0.36 world-knowledge claims per answer against the backbone's 0.37 (shares 11.9 vs. 12.4%), and their audited support rate rises from 93.6% to 95.9% (the complement of the hUCR values in the table). All rates in Table B.1 use the strict criterion; the type shares weight them to the pooled hUCR of Table 2 (11.0 / 4.6%) and to the ledger's NULL rate (5.9%).

**A fourth grader outside the Qwen family.** Because J-ind is fine-tuned from Qwen2.5-VL-72B, a Gemma-3-27B classifier fine-tuned on the same 12k-claim set grades the headline systems as a check on family preference (strict criterion): backbone 12.1%, on-policy grounder-DPO 8.0%, on-policy

Table B.1: UCR by claim type on the 520-prompt audit (human grader), the ledger's NULL rate by type, and the share of claims per type (backbone / ledger).

| Claim type | share of claims (%) | hUCR backbone (%) | hUCR ledger (%) | ledger NULL rate (%) |
|---|---|---|---|---|
| Perception | 41.2 / 40.6 | 14.2 | 5.6 | 5.8 |
| OCR / numeric | 27.5 / 28.1 | 9.6 | 3.4 | 4.4 |
| Reasoning step | 18.9 / 19.4 | 8.6 | 4.7 | 6.5 |
| World knowledge | 12.4 / 11.9 | 6.4 | 4.1 | 9.2 |

GRPO $7.8\%$, grounder control $11.5\%$, ledger $5.4\%$ (Kendall $\tau = 0.89$ with J-ind over all nineteen systems).

**Notation.** Thresholds: $\tau_{\text{sup}}$ (support threshold, selected as $\hat{\tau}_{\text{sup}}$ by risk control), $\tau_{\varnothing}$ (null-dominance threshold, $0.50$), $\tau_{\text{cyc}}$ (cycle threshold, $0.05$), $\tau_{\text{contra}}$ (contradiction share, $0.25$), $\tau_{\text{ul}}$ (unlikelihood onset, $0.30$); the strict support criterion is $s < \frac{1}{2}$ and uses no $\tau$. Risk control: $\vartheta$ (target false-SUPPORTED rate, $0.03$), $\delta$ (failure probability, $0.05$; confidence level $1 - \delta = 0.95$). Solver: $\varepsilon$ (entropic regularisation, $0.05$), $\epsilon_b$ (composition guard, $10^{-6}$), $T$ (iterations, $15$), $K$ (ANN candidates, $8$), $K_{\text{prior}}$ (protected prior edges, $4$), $K_{\text{eff}} = K + K_{\text{prior}}$; degree penalty: $\varsigma_s$ (soft-degree scale, $0.005$), $K_c$ / $K_r$ (column / row caps, $8$ / $2$), $\varsigma$ (hard-count threshold for reporting, $10^{-3}$); emission: $\zeta$ (funding-set share, $0.10$), $\nu_0$ (constant null cost, $0.62$); loss weights $(\alpha, \beta, \gamma, \eta)$ and $(\lambda_{\varnothing}, \lambda_g, \lambda_s)$. No symbol is used with two meanings.

**Metric definitions.** UCR is the percentage of parsed atomic claims of the answer *as emitted* that are unsupported under the strict criterion ($s < \frac{1}{2}$: any typed slot scored 0) and the stated grader (verifier, frozen auxiliary evaluator J-ind, or human panel); Table D.5 reports the lenient slot-average criterion for comparison; NULL rows are flags and are never removed before scoring, so a certificate cannot improve UCR by selective reporting. Certificate $F_1$ (Cert.$F_1$) is the funding-set $F_1$ of the certificate as emitted: for every claim the set of evidence identifiers cited by its row (the atoms at share $\geq \zeta$; a NULL row cites no evidence and contributes an empty set) is compared with the adjudicated evidence identifiers, and precision and recall are pooled over claims. Cert.$F_1$ therefore depends on the NULL rule ($\tau_{\varnothing}$ and the contradiction flag) but not on $\hat{\tau}_{\text{sup}}$ or $\tau_{\text{cyc}}$, which only exchange SUPPORTED and WEAK rows. The *raw* funding-set $F_1$, computed from the funding sets of all claims before any verdict is assigned, involves no threshold and is the quantity maximised to fit $\nu_0$ (Appendix C); on the ledger it is $0.4$ above Cert.$F_1$. Verdict quality is reported separately in Table K.24. ECE is the expected calibration error of $\hat{s}$ in ten equal-mass bins and carries a bootstrap standard error of $0.3$. *Claim coverage* is the fraction of parsed claims stamped SUPPORTED or WEAK; *answer coverage* is the fraction of prompts that receive a non-refusal answer (Table 2); the matched-coverage controls are matched on claim coverage. For systems that emit no certificate, Cert.$F_1$, ECE, and claim coverage are computed by attaching our frozen post-hoc grounder (trained on the 6.2k annotations and frozen before any comparison) to their answers, and are marked $^{\dagger}$; they measure that grounder on those answers, not the systems themselves. ME-AUC is the area under the abstention curve when decisive evidence is removed. Hall.Acc is HallusionBench question-pair accuracy (qAcc, the harness default); per-question aAcc and figure-level fAcc are in Table K.26. POPE $F_1$ is accompanied by the yes-ratio in the same table. Gap is the gradient-carrying $\Delta^{\nabla}_{\text{led}}$ of Eq. (2) ($0.027$ at convergence for the full ledger: $\Delta_{\text{cyc}}$ $0.019$, $\lambda_g \Gamma$ $0.003$, $\lambda_s \Omega_s$ $0.005$; the overflow part of $\Gamma$ is $0.002$); the gradient-free $\Delta_{\text{null}}$ is reported separately ($0.081$ for the full ledger, of which $0.041$ from $\Pi^{\text{EC}}$ and $0.040$ from $\Pi^{\text{EA}}$), and the diagnostic of Figure A.4c is $\Delta_{\text{cyc}} + \Delta_{\text{null}}$ ($0.019 + 0.081 \approx 0.10$ on clean data). Rows without a direct witness or without transport have no $\Delta_{\text{cyc}}$ and show "–". Latency is wall-clock inference time relative to the unmodified backbone served locally with the same engine; the API model has no comparable figure.

**The auxiliary evaluator J-ind.** J-ind is a claim-support classifier fine-tuned from Qwen2.5-VL-72B on a 12,000-claim annotation set produced by a separate annotation team before this work and frozen before ledger training began. It shares no prompts, images, or annotators with the 6.2k annotations used to train the ledger and the controls, and it was never updated during this work.

**Weighting conventions.** All claim-level rates are claim-micro (pooled) unless labelled "answer-macro". Calibration and audit quantities are estimated with claim-micro weighting within their stated

populations; the audit's answer-macro rates ($10.6/7.0/9.9/4.3\%$ for backbone / grounder-DPO / control / ledger) are given for comparison and agree with the micro rates to within $0.4$ pp.

**Claim specificity and answer quality.** Equal claim counts do not exclude a drift towards vaguer claims, which the strict criterion would reward. Three measures on the 520-prompt audit show none: the mean number of filled typed slots per claim is $2.71$ (ledger) against $2.70$ (backbone); the share of claims whose head noun is a WordNet hypernym of a noun the backbone used for the same prompt ("an animal" for "a golden retriever") is $4.1\%$ against $3.9\%$; the share of claims carrying a numeric, date, or count value is $28.1\%$ against $27.5\%$. Blinded answer-quality ratings (1–5, two raters, 200 prompts) are $4.21$ (ledger) and $4.19$ (backbone). The ledger does say less about what its tools cannot see: relative to the backbone it emits $0.9$ pp fewer correct claims about objects smaller than $32^2$ pixels, $1.2$ pp fewer about object classes outside the COCO vocabulary of the detector, and $2.4$ pp fewer about text in non-Latin scripts (shares of all emitted claims on the audit sample), and its NULL over-flag rate on these categories is $12.4\%$ against $5.9\%$ overall. This is the cost of weighting unlikelihood by what the tools atomise.

**Claims on multiple-choice and grounding benchmarks.** The neutral system prompt asks for concise factual statements, so on MMMU, MMMU-Pro, MMBench, MMStar, AI2D, EgoSchema, and HallusionBench a typical answer is one or two justifying statements followed by the option letter or yes/no token, which the harness extracts with its default regular expressions; accuracy with and without the justification differs by at most $0.2$ on every multiple-choice benchmark, and the claims are parsed from the statements. On RefCOCOg the answer is a bounding box followed by one descriptive sentence, and the claims come from that sentence; the box is scored by the harness. On POPE the answer is yes/no followed by one statement. Median parsed claims per answer are 3 on every benchmark family except POPE (2).

**Annotator pools.** Three disjoint pools of trained, compensated annotators were used, all working from the same annotation manual (version 3): pool A (nine annotators) produced the 6.2k training annotations and the 4,000-claim slice on which the calibrated-verifier grader was trained and calibrated; pool B (five annotators) produced the 2,000-claim risk-control set of every system; pool C (six annotators and one adjudicator) produced the 520-prompt audit, the manual atomization, the seeded-error study, and the InternVL audit. The 12k-claim J-ind set was produced by a separate team before this work. No annotator belongs to two pools; the shared manual is a shared convention, and the label-noise analysis of Appendix D bounds its effect on the guarantee.

## C REPRODUCIBILITY MANIFEST

**Backbones.** `Qwen/Qwen3-VL-235B-A22B-Instruct` (sparse MoE, 22B active), frozen. All backbone scores are our own re-evaluation under VLMEvalKit v0.5.4, the neutral system prompt, greedy decoding, and the evaluation slices of Table C.2; MMMU-Pro is the standard ten-option split with a direct-answer prompt and no chain of thought. These settings differ from those of the model card (prompt, answer format, and, for MMMU-Pro, the split and option format), so the two sets of numbers are not directly comparable; every comparison reported here is between systems evaluated under the same settings. `OpenGVLab/InternVL3_5-38B` (dense LLM, ViT-6B tower), frozen; LoRA adapters trained. Every baseline checkpoint is listed in Table C.3.

**Evidence tools.** Detector: Grounding DINO (`IDEA-Research/grounding-dino-base`, box threshold $0.30$, at most 40 boxes); OCR: PaddleOCR 2.7 (PP-OCRv4, line level); chart parser: DePlot (`google/deplot`) producing series, axis, and point atoms; video: 1 fps frames with SigLIP shot-boundary merging (at most 32 frames). The same tools and versions serve every system that uses evidence atoms (ledger, grounder controls, Control A).

**Atomizers and embeddings.** The ledger's trainable claim atomizer $h^C_\psi$ is a text-side model: a six-layer, 384-dimensional transformer encoder (MiniLM-sized, 21M parameters including its linear map to $\mathbb{R}^d$) over the decoded answer *text*, which tags claim spans, classifies the (entity, attribute, value) slots, and emits the claim embedding $z^C_j \in \mathbb{R}^d$ by mean-pooling the span and applying the linear map. It reads tokens and character offsets, never hidden states, which is why it transfers frozen

across backbones (Section 5.5). The evidence projector $h^E_\psi$ (21M) maps each tool atom (its detector, OCR, or chart-parser embedding concatenated with the frozen vision-encoder features of its region) to $z^E_i \in \mathbb{R}^d$, $d = 1{,}024$. Answer-unit embeddings $z^A_k$ are a linear map (part of $h^E_\psi$'s parameter count) of the mean-pooled final decoder states over the unit's tokens, so that gradients on $z^A$ reach the LoRA adapters; on InternVL only this linear map's input dimension changes ($5{,}120$ instead of $4{,}096$). The verifier head $v_\omega$ (11M) is a two-layer cross-attention block in which the claim embedding attends to the candidate atoms' embeddings, concatenated with the column $\Pi^{\mathrm{EC}}_{\cdot j}$, followed by an MLP that outputs the raw support logit $u_j$; $u_j$ is trained by $\mathcal{L}_{\mathrm{ver}}$ and $\mathcal{L}_{\mathrm{cal}}$ and becomes the first feature of $\varphi_j$ in Eq. (4), whose logistic and isotonic maps are fit post hoc and frozen. Nothing else produces $\hat{s}_j$. The *evaluation* atomizer that parses every system's answers for UCR is a different model: a text-only Qwen2.5-7B sequence-to-sequence decomposer trained before this work on a pre-existing 1,500-answer decomposition set (not part of the 46k records), frozen, and measured on the disjoint 400-prompt (1,200-claim) parser-eval set (recall $93.6\%$, precision $94.2\%$ against the union of two human decompositions; $92.8\%$ / $93.7\%$ on the 250-prompt manual-atomization audit). It shares no architecture, initialisation, training data, or hidden states with $h^C_\psi$, which reduces the risk that biases in the ledger's atomizer directly determine its evaluation scores.

**Trainable footprint.** Evidence projector $h^E_\psi$ including the answer-unit map (21M), claim atomizer $h^C_\psi$ (21M), verifier $v_\omega$ (11M), and LoRA rank 16 ($\alpha$=32, dropout $0.05$) on the query and value projections of all 94 decoder blocks: with hidden size 4,096, 64 query heads and 4 key/value heads of dimension 128, one block contributes $16 \times (4{,}096 + 8{,}192) + 16 \times (4{,}096 + 512) = 270{,}336$ parameters, and 94 blocks $25.4$M; total $78$M trainable. Because adapters sit in every block, the backward pass traverses the whole network, which is the assumption of the compute derivation in Appendix C.1.

**Optimiser and schedule.** AdamW, $\beta_1$=0.9, $\beta_2$=0.95, weight decay $0.1$; peak LR $5 \times 10^{-5}$ (LoRA) and $2 \times 10^{-4}$ ($h_\psi$, $v_\omega$); cosine schedule, 500-step warm-up, 10,550 optimiser steps; micro-batch 4 per device on 16 devices, gradient accumulation 8 (effective 512 sequences); three epochs over $1.8$M pairs $= 5.4$M sequences; mean sequence length 907 tokens including visual tokens and the teacher-forced greedy draft ($\approx 180$ tokens) on which $\mathcal{L}_{\mathrm{ul}}$ and the ledger are computed, i.e. $4.9$B ledger-training tokens per run; bf16; gradient clip $1.0$.

**Ledger hyper-parameters and how they were chosen.** $K$=8, $K_{\mathrm{prior}}$=4, $T$=15, entropic $\varepsilon$=0.05, $\zeta$=0.10 (normalised share), soft-degree scale $\varsigma_s$=0.005 with caps $K_c$=8, $K_r$=2, $(\alpha, \beta, \gamma, \eta) = (1.0, 0.5, 0.5, 0.25)$, $\lambda_\varnothing$=1, $\lambda_g$=0.5, $\lambda_s$=0.1, $\tau_{\mathrm{ul}}$=0.30, thresholds $\tau_\varnothing$=0.50, $\tau_{\mathrm{cyc}}$=0.05, $\tau_{\mathrm{contra}}$=0.25, and $\hat{\tau}_{\mathrm{sup}}$=0.50 as selected by risk control at $\vartheta$=0.03, $\delta$=0.05. Loss weights, $K$, $T$, $\varepsilon$, $\varsigma_s$, $K_c$, $K_r$, and $\tau_{\mathrm{ul}}$ were chosen on a Qwen3-VL-2B development sweep (verifier UCR and Cert.$F_1$ on a 1,000-prompt development slice drawn from training-split prompts) and then fixed for every backbone; $\tau_\varnothing$, $\tau_{\mathrm{cyc}}$, $\tau_{\mathrm{contra}}$, and $\zeta$ were chosen on the verifier-fit split by maximising verdict agreement with adjudication at fixed coverage; none was tuned on evaluation, calibration-test, or audit data. The constant null cost $\nu_0$=0.62 was fit once on the scorer-fit subset by maximising the raw funding-set $F_1$ (funding identifiers before any verdict is assigned; no threshold enters) of the ledger modules after their 500-step warm-up on the answer-only LoRA drafts, and frozen; re-fitting it at the end of training changes it by $0.02$ and Cert.$F_1$ by $0.1$; capacity weights $w_{\mathrm{mod}} \in [0.8, 1.2]$ per modality, fit once and frozen; hard-count threshold $\varsigma$=$10^{-3}$ (Eq. (7)).

**Data.** 1.8M instruction pairs from LLaVA-OneVision-Mix, ShareGPT4V, ChartQA-train, DocVQA-train, AI2D-train, A-OKVQA-train, RefCOCOg-train; 46,000 claim-level records: 6,200 human-annotated claims (5,000 verifier-fit claims with gold spans/boxes/segments, and a parser-evaluation set of 400 prompts containing 1,200 claims) and 39,800 swap pairs constructed automatically from the verifier-fit claims by exchanging their decisive atoms within a benchmark, validated on a 1,000-pair human check ($96.3\%$ valid). We refer to the human-annotated part as the $6.2$k annotations and to the whole as the $46$k records. The 2,000 risk-control claims per system (Appendix D) are annotated after training and are not part of this set.

**Prompt and inputs.** One neutral system prompt ("You are a careful multimodal assistant. Answer the question with concise factual statements.") precedes the unmodified benchmark prompt for every same-backbone system, baselines and controls included. The decoder receives the backbone's visual

tokens and this text only: evidence atoms (detector boxes, OCR strings, chart elements) are consumed by the ledger side alone and are never inserted into the ledger's decoder input. The tool-in-context baselines are the deliberate exception: they receive the same tools' outputs as additional text in the user turn; all systems otherwise use the same benchmark inputs and base prompt.

**Decoding.** Greedy, temperature 0, max 512 tokens; the certificate is emitted once over the finished answer and never edits it.

**Harness.** VLMEvalKit v0.5.4 with default per-benchmark protocols; evaluation sizes in Table C.2.

Table C.2: Per-task evaluation sizes and splits; the same slice is used for every system. Risk-control prompts are drawn from each benchmark's training or dev split where one exists and never from the evaluated split; for the four benchmarks without one (POPE, HallusionBench, MMStar, OCRBench) a fixed 10% carve-out is removed from the evaluation set for every system, and the sizes listed are after the carve-out.

| Dataset | Split | $n_{\text{eval}}$ | Dataset | Split | $n_{\text{eval}}$ |
|---|---|---|---|---|---|
| MMMU | val (full) | 900 | POPE | adversarial (after 10% carve-out) | 2700 |
| MMMU-Pro | test (overall) | 1730 | HallusionBench | val (after carve-out) | 856 |
| MMBench-DEV-EN | dev | 4377 | VideoMME | sub (short+med+long) | 2700 |
| MMStar | test (after carve-out) | 1350 | EgoSchema | full | 5063 |
| DocVQA | val | 5349 | RefCOCOg | val | 2573 |
| TextVQA | val | 5000 | Risk-control set (claims, post-training) | one claim per drawn answer | 2000 |
| ChartQA | test (avg human/aug) | 2500 | OCRBench | full (after carve-out) | 900 |
| AI2D | test | 3088 | Verifier-fit split (claims) | training annotations | 5000 |

Table C.3: Checkpoints and API versions behind every system name. Frozen systems are single greedy runs; fine-tuning baselines use the authors' released code and data on our backbone.

| System name | Checkpoint / version | Notes |
|---|---|---|
| GPT-4o | `gpt-4o-2024-11-20` (API) | queried Nov 2025; temperature 0; no latency comparison |
| Llama-4-Maverick-17B-128E | `meta-llama/Llama-4-Maverick-17B-128E-Instruct` | natively multimodal MoE, 17B active |
| InternVL3.5-241B-A28B | `OpenGVLab/InternVL3_5-241B-A28B` | largest released InternVL3.5 |
| Qwen3-VL-235B-A22B | `Qwen/Qwen3-VL-235B-A22B-Instruct` | backbone of all same-backbone systems |
| Scaling sweep | `Qwen/Qwen3-VL-{2B,4B,8B,30B-A3B,32B}-Instruct` | released sizes; one run per cell |
| InternVL replication | `OpenGVLab/InternVL3_5-38B` | dense; three seeds |
| DoLa / VCD / OPERA | authors' decoding code on the backbone | VCD noise seeds $\{17, 29, 41, 53, 67\}$ |
| Woodpecker | authors' pipeline; GPT-4o-mini as the corrector | frozen pipeline, single run |
| LURE / HA-DPO / POVID / RLAIF-V | authors' recipes and released preference data, LoRA rank 16 on the backbone | three seeds each; RLAIF-V at rank 64: UCR 7.8 vs. 7.9 |
| On-policy grounder-DPO | preference pairs from the model's own drafts scored by our grounder, DPO at matched compute (695 active H100-hours per seed, including the drafts) | three seeds; reference baseline in the audit and stress tests |
| On-policy GRPO | group-relative policy optimisation with a claim-level reward ($+1$ per claim the frozen grounder marks supported, $-1$ per claim it marks unsupported, $-0.2$ per abstention), 4 rollouts per prompt, KL coefficient 0.02, at matched compute | three seeds; within 0.2 pp of grounder-DPO on every unsupported-claim metric |
| Tool outputs in context | the OCR lines, detector boxes with labels, and chart-parser series of the same tools, serialised as text (at most 512 tokens) and prepended to the user turn of the frozen backbone | single greedy run; the same serialisation is used for the DPO-tuned variant |
| Tool outputs in context + grounder-DPO | the serialisation above with on-policy grounder-DPO trained on drafts produced under it, at matched compute | three seeds; strongest baseline on every unsupported-claim metric in Table 1 |
| J-ind | Qwen2.5-VL-72B fine-tuned on a disjoint 12k-claim set | frozen before this work |

**Certificate schema.** Each certificate is a JSON object with the answer text, the atom list (identifier, modality, location, capacity), and one row per claim: claim text and span offsets, typed slots, funding identifiers with shares, $\bar{c}_j$, $\bar{\delta}_j$, $n_j^{\text{alloc}}$, $r_j$, $u_j$, $\hat{s}_j$, the contradiction flag with the offending atom identifier, and the verdict. Every certificate figure and gallery row is rendered from these objects, and every table is derived from per-run evaluation files, one per run and benchmark.

**The counterfactual swap term.** For a claim $c_j$ with decisive atom $i_j^\star = \arg\max_i \Pi_{ij}^{\text{EC}}$ (selection stop-gradient), a swap replaces the atom *embedding* $z_{i_j^\star}^E$ by the embedding $z_{i'_j}^E$ of an atom of the same modality drawn from another example whose entity label (detector class, OCR string, or chart series

name) does not match any slot of $c_j$; the input image and the decoder forward pass are unchanged (atoms never enter the decoder), so $z^A$ is the original one, and the three plans are re-solved on the swapped atom set. With $\bar{\Pi}^{CA}_{jk} = \Pi^{\mathrm{CA}}_{jk} / \sum_{k'} \Pi^{\mathrm{CA}}_{jk'}$,

$$\mathcal{L}_{\mathrm{cf}} = \frac{1}{|J_{\mathrm{cf}}|} \sum_{j \in J_{\mathrm{cf}}} \Big[ \frac{\tilde{\Pi}^{EC}_{i'_j j}}{b_j} + \sum_k \bar{\Pi}^{CA}_{jk} \frac{\tilde{\Pi}^{EA}_{i'_j k}}{q_k} \Big], \tag{6}$$

the budget-normalised funding that the swapped-in atom receives on the composed route and on the direct route; it is zero exactly when the funding set has moved away from the swapped position. The first term back-propagates to the projector and the atomizer through $\tilde{D}^{EC}$, the second to the adapters through $z^A$ in $\tilde{D}^{EA}$, which is why the additive row "LoRA $+\mathcal{L}_{\mathrm{cf}}$ only" changes the generator. $J_{\mathrm{cf}}$ contains one claim per sequence for half of the sequences of every batch (the compute accounting of Appendix C.1 charges one counterfactual solve per such sequence): the pre-constructed pair from the 39,800 swap records when the sequence has one, otherwise a swap drawn on the fly from another sequence of the same modality in the batch.

**Cost terms, capacities, and protected priors.** $D^{EC}_{ij} = 1 - \cos(z^E_i, z^C_j) + \rho_{\mathrm{mod}}(i) + \rho_{\mathrm{geo}}(i,j)$ with $\rho_{\mathrm{mod}}(i) = \kappa_{\mathrm{mod}}(1 - \mathrm{conf}_i)$, $\kappa_{\mathrm{mod}} = 0.3$, where $\mathrm{conf}_i$ is the tool's own confidence (detector score, OCR line confidence, chart-parser confidence; 1 for video frames), and $\rho_{\mathrm{geo}}(i,j) = \kappa_{\mathrm{geo}}\, d(i,j)$, $\kappa_{\mathrm{geo}} = 0.2$, where $d(i,j) \in [0,1]$ is the normalised distance between the atom's location and the location of the claim's grounded entity when the claim has one (a referring box or a cited region) and 0 otherwise; $D^{CA}$ and $D^{EA}$ use the same form with $z^A$ in place of $z^C$ and no geometric term for $D^{CA}$. Capacities are $u_i = \min(1, \mathrm{conf}_i\, w_{\mathrm{mod}(i)})$ with the per-modality weights $w_{\mathrm{mod}} \in [0.8, 1.2]$ above. The candidate set of a column is the top-$K$ atoms by cost together with up to $K_{\mathrm{prior}} = 4$ *protected prior* atoms whose label or text lexically matches the claim's entity slot (case-folded token overlap $\geq 0.5$); protected atoms are always in the support regardless of their cost, and the union has at most $K_{\mathrm{eff}} = 12$ entries.

**The InfoNCE control.** Control C and the two InfoNCE additive rows replace $\Delta_{\mathrm{cyc}}$ by a symmetric InfoNCE loss between each answer-unit state $z^A_k$ and its matched atoms with temperature 0.07 and in-batch negatives; the matched atoms are those in the funding set of the unit under $\Pi^{\mathrm{EA}}$ at the current step for Control C, and under the frozen reference ledger for the additive rows, so the control differs from the ledger in the form of the alignment term and in nothing else.

**Which sequences receive which loss.** Every sequence contributes $\mathcal{L}_{\mathrm{gold}}$ on its gold answer; every sequence also receives an on-line greedy draft, on which $\mathcal{L}_{\mathrm{ul}}$ and $\Delta^{\nabla}_{\mathrm{led}}$ are computed without labels; $\mathcal{L}_{\mathrm{cf}}$ is evaluated on half of the sequences as described above; $\mathcal{L}_{\mathrm{ver}}$ and $\mathcal{L}_{\mathrm{cal}}$ are evaluated only on sequences whose claims carry gold rationales (the 3,000 head-training claims of the verifier-fit split described below), which are oversampled so that every batch contains some. The on-policy grounder-DPO and GRPO baselines start from the same answer-only LoRA warm start on the same 1.8M pairs and then optimise their preference or reward objective on their own drafts at matched compute, so the comparison isolates the grounding objective and not the instruction data.

**Splits used for fitting the scorer.** The 5,000 verifier-fit claims are partitioned once: 3,000 train $\mathcal{L}_{\mathrm{ver}}$ and $\mathcal{L}_{\mathrm{cal}}$, and the remaining 2,000 (the scorer-fit subset, never seen by $v_\omega$) fit the logistic map $w$, the isotonic map $g_{\mathrm{iso}}$, $\nu_0$, and the thresholds $\tau_\varnothing$, $\tau_{\mathrm{cyc}}$, $\tau_{\mathrm{contra}}$, and $\zeta$; wherever this manifest says "verifier-fit split" for one of these post-hoc fits, the scorer-fit subset is meant. The learn-then-test threshold uses the separate 2,000-claim risk-control set, and the guarantee does not depend on this partition.

**The audited seed.** The human audit uses seed 17 for every system. Under the verifier grader seed 17 is the ledger's weakest seed (5.23% against a five-seed mean of 4.91%) and the answer-only LoRA generator's strongest (10.41% against 10.84%); the per-seed values of every audited trained system are listed in Table G.15, so the position of the audited seed relative to each system's mean can be read off directly.

**Tuning budgets.** Every trained system received the same development budget: twelve configurations on Qwen3-VL-2B, scored by verifier UCR and Cert.$F_1$ on a 1,000-prompt development slice drawn from training-split prompts, followed by a single run of the selected configuration at scale per seed. For the ledger the twelve configurations spanned the loss weights, $K$, $T$, $\varepsilon$, and $\tau_{\text{ul}}$; for grounder-DPO, $\beta \in \{0.05, 0.1, 0.2\}$, learning rate $\in \{10^{-6}, 5 \times 10^{-6}, 10^{-5}\}$, and pair-construction thresholds; for GRPO, the KL coefficient $\in \{0.01, 0.02, 0.05\}$, rollouts $\in \{4, 8\}$, and the abstention penalty; for the released-recipe baselines, the authors' defaults plus learning rate and LoRA rank. The 39,800 swap pairs are part of the 46k records available to every control trained on our annotations, including the grounder and the claim classifier, which use them as hard negatives.

**An 8B reference configuration.** To make the method checkable without a 235B backbone, the same recipe on `Qwen/Qwen3-VL-8B-Instruct` (three seeds $\{17, 29, 41\}$, identical hyper-parameters except effective batch 256 and 21,100 steps, 253 active H100-hours per seed including drafts and counterfactual re-solves) gives strict verifier UCR 11.0 $\rightarrow$ 6.9% (answer-only LoRA 10.6%), Cert.$F_1$ 81.3, claim coverage 92.0%, HallusionBench qAcc +7.9, POPE $F_1$ +2.6, and $\text{Avg}_{14}$ +1.4, with seed SDs of 0.19 pp (UCR) and 0.31 ($F_1$); the 8B/4.9B cell of Table G.21 is its seed-17 run.

### C.1 Compute accounting

The unit throughout is *active device-hours*, $\sum_{\text{jobs}}(\text{devices} \times \text{active elapsed hours})$, excluding queueing and preemption. Every training figure is derived from processed tokens and active parameters (caption of Table C.4) and cross-checked against job logs. Each main run used 16 H100s for 21.6 wall-clock hours for the LoRA pass (adapters in all 94 blocks, so the backward pass traverses the whole network; the teacher-forced draft is part of every sequence) plus a separate 8×H100 vLLM inference service for 43.5 hours that produced the on-line drafts (weights synchronised every 200 optimiser steps; logged throughput 6,200 tokens/s aggregate at batch 256); the five main runs spanned 6.8 days including two restarts. Peak allocation was 40 H100s.

Table C.4: Additive compute accounting (active H100-hours). Training cost is derived from processed tokens and active parameters: a frozen backbone with LoRA adapters in every block needs a forward pass and an activation-gradient backward pass through the whole network but no weight gradients for the frozen parameters, $\approx 4 \times P_{\text{active}} \times$tokens FLOPs, evaluated at 35% MFU on H100 (989 TFLOP/s dense bf16); the token count per run (4.9B) includes the teacher-forced greedy draft on which the ledger and $\mathcal{L}_{\text{ul}}$ are computed, so no separate pass is needed for it; e.g. $4 \times 22\text{B} \times 4.9\text{B}/(0.35 \times 989\,\text{TFLOP/s}) \approx 346$ hours per seed at 235B-A22B for the LoRA pass; on-line greedy drafts ($\approx 180$ tokens per sequence, 0.97B tokens per run) are produced by a separate 8×H100 vLLM service whose weights are synchronised every 200 optimiser steps, at a measured aggregate throughput of 6,200 tokens/s (decoding is memory-bound, so no MFU assumption is used), and the counterfactual swap re-runs only the sparse solver on cached atom, claim, and answer-unit embeddings for half of the sequences (no decoder pass; the swapped atom never enters the decoder), costed at 0.5% of the LoRA pass: they add 348 and 2 hours, giving 695 per seed. The on-policy grounder-DPO and GRPO baselines are costed at the same 695 hours per seed as the main configuration, drafts or rollouts included. Measured job accounting agreed with these derived figures to within 7%.

| Run group | Runs | Tokens / run | Active H100-hours |
|---|---|---|---|
| Qwen3 main configuration (per seed: LoRA pass 346 + on-line drafts 348 + counterfactual re-solves 2) | 5 seeds | 4.9B | 3,477 |
| Attribution controls A, B1, B2, C (Table 3) | 4 × 3 seeds | 4.9B | 7,511 |
| On-policy grounder-DPO and on-policy GRPO baselines (matched: 695 h per seed each) | 2 × 3 seeds | 4.9B | 4,173 |
| Portability rows (grounder-DPO $+\Delta_{\text{cyc}}$, $+\Delta_{\text{cyc}} + \mathcal{L}_{\text{cf}}$, $+\mathcal{L}_{\text{ul}}$, all three) | 4 × 3 seeds | 4.9B | 8,345 |
| Calibration fitting (logistic + isotonic + LTT threshold) | 5 | forward only | 28 |
| Evaluation, all systems, 14 benchmarks + long-form benchmarks + RePOPE + audits | 21 systems | – | 645 |
| Ablations and additive controls (Table G.11) | 24 retrained configs × 3 seeds | 4.9B | 48,068 |
| $K$ and $T$ sensitivity (Table G.17) | 6 configs × 3 seeds | 4.9B | 12,142 |
| Scaling sweep (Table G.21), main cell excluded | 35 runs | 0.15–4.9B | 3,679 |
| Scaling-sweep replicate seeds (four cells × 2 extra seeds) | 8 runs | 0.15–4.9B | 2,321 |
| Matched-data controls on the 6.2k annotations | 3 × 5 seeds | ≤4.9B | 3,289 |
| Fine-tuning baselines (LURE, HA-DPO, POVID, RLAIF-V) | 4 × 3 seeds | own recipes | 1,800 |
| Development and failed runs | 38 | – | 4,104 |
| InternVL3.5-38B matched answer-only LoRA | 3 seeds | 4.9B | 1,685 |
| InternVL3.5-38B full ledger retraining | 3 seeds | 4.9B | 3,604 |
| InternVL3.5-38B transfer ablation | 3 seeds | 1.9B | 1,369 |
| InternVL3.5-38B evaluation | 3 systems | – | 60 |
| **Project total** | 247 runs | | **≈106,300** |

Per-seed training cost including drafts and counterfactual re-solves is $\approx 63/1{,}201/695$ active H100-hours at Qwen3-VL-2B / InternVL3.5-38B / Qwen3-VL-235B-A22B; the dense 38B model is more expensive than the 22B-active MoE at equal tokens. Adaptation is one-time; the transfer path (1.9B

tokens, projector and adapters only) costs $38\%$ of full retraining; inference overhead thereafter is $1.18\times$.

## D CALIBRATION MEASUREMENTS AND AUDIT PROTOCOL

**Latent support $s$.** Annotators decompose each claim into typed slots (entity, attribute, value) and score each slot $1/0.5/0$ for entailment by the gold rationale spans, boxes, or temporal segments; $s$ is the adjudicated *minimum* over slots and is defined for every parsed claim. "Unsupported" means $s < \frac{1}{2}$ throughout, i.e. any slot scored 0 (strict criterion). Under the lenient slot-average criterion a claim with a correct entity and attribute but a wrong value scores $s = \frac{2}{3}$ and passes; Table D.5 reports both criteria on the same adjudications.

Table D.5: Strict (minimum over slots) versus lenient (slot average) criterion on the same 520-prompt adjudications. The strict criterion is used throughout; the gap between criteria is largest for OCR/numeric claims, where value errors on otherwise correct claims are common.

| System | hUCR strict (%) | hUCR lenient (%) | value-only errors (% of claims) | ahUCR strict / lenient (%) |
|---|---|---|---|---|
| Qwen3-VL-235B | 11.0 | 8.6 | 2.4 | 11.0 / 8.6 |
| + on-policy grounder-DPO | 7.2 | 5.7 | 1.5 | 7.2 / 5.7 |
| Answer-only LoRA + grounder @ 94.1 | 10.3 | 8.1 | 2.2 | 8.6 / 6.7 |
| LEDGER-VLM | 4.6 | 3.6 | 1.0 | 2.5 / 2.0 |

**Score fitting and risk-control claims.** The scorer-fit subset of the verifier-fit split (2,000 claims never seen by $v_\omega$; Appendix C) fits $w$ and $g_{\text{iso}}$ of Eq. (4). The threshold $\hat{\tau}_{\text{sup}}$ is then selected by the learn-then-test procedure of Theorem 1 on a 2,000-claim risk-control set built after training: 2,000 draws, each of which samples a prompt from a held-out slice of the fourteen benchmarks *with replacement* and with probability proportional to its answer's number of parsed claims, decodes (or reuses) the final model's answer, and keeps one of its parsed claims uniformly at random; 1,871 distinct prompts were drawn and 121 draws repeated a prompt. The draws are i.i.d. from $\mathcal{D}_{\text{claim}}$, so the counts are exact binomials (a repeated prompt yields the same answer but an independent claim draw, and the two draws are independent copies of the same random variable, not dependent claims of one answer); the set shares no prompt or image with training, with the 46k records, or with the audit; and each control system has its own set built from its own outputs. The claims are annotated by pool B of Appendix B, which is disjoint from the audit pool. $\mathcal{D}_{\text{claim}}$ is therefore the uniform distribution over the pooled parsed claims of the deployed system on the held-out prompt pool; the guarantee of Theorem 1 concerns deployment claims exchangeable with that pool, and its extension to a new prompt stream is the distributional assumption tested in Table D.8, not a consequence of the theorem. At $\vartheta = 0.03$, $\delta = 0.05$ and a 40-point threshold grid ($0.025$ spacing, tested from $0.975$ downwards) the procedure selects $\hat{\tau}_{\text{sup}} = 0.500$ with $22/1{,}296$ false-SUPPORTED calibration claims ($1.7\%$; one-sided binomial $p = \Pr\{\text{Bin}(1{,}296, 0.03) \le 22\} = 0.002$). The fixed sequence is monotone in the counts, listed as threshold, false-SUPPORTED count over SUPPORTED count, and $p$-value: at $0.900$, 6 of 812 and $p < 10^{-4}$; at $0.800$, 10 of 1,002 and $p < 10^{-4}$; at $0.700$, 14 of 1,121 and $p = 10^{-4}$; at $0.600$, 17 of 1,204 and $p = 3 \times 10^{-4}$; at $0.550$, 20 of 1,252 and $p = 0.001$; at $0.525$, 21 of 1,275 and $p = 0.002$; and at $0.500$, 22 of 1,296 and $p = 0.002$. Every grid point above $0.500$ has $p \le 0.002$; the next point, $0.475$, has $30/1{,}320$ and $p = 0.067$, so testing stops there. The coincidence of the selected value with the round number $0.500$ is a property of the grid; with a 200-point grid the procedure selects $0.495$ and coverage changes by $0.3$ pp. If the first grid point were not rejected, no threshold would be certified, every non-NULL row would be stamped WEAK, and certified coverage would be zero until recalibration; this did not occur. The threshold is selected separately for each training seed on that seed's own calibration draw; all five seeds selected $0.500$, and the reported $\hat{\tau}_{\text{sup}}$, counts, and $p$-value are those of seed 17, whose test claims are audited. On the audited test claims the false-SUPPORTED rate is $24/1{,}031 = 2.3\%$ (prompt-clustered bootstrap $[1.5, 3.4]$). Reliability bins are in Table D.7.

**Mapping certificate rows to evaluation claims.** Certificate rows are produced on the ledger's atomization $h^C_\psi$ and every estimand is computed on the evaluation atomizer's claims. The two agree on $95.1\%$ of spans (character-overlap Jaccard $\ge 0.7$), and an agreed claim inherits its row's verdict.

For the remaining $4.9\%$: a split (one ledger row covering two evaluation claims, $2.6\%$) gives both claims the row's verdict; a merge (two rows covering one evaluation claim, $1.8\%$) gives the claim the more conservative verdict (NULL over WEAK over SUPPORTED); an unmatched evaluation claim ($0.5\%$) receives no verdict and is counted as *not asserted*, so it lowers coverage and cannot lower the false-SUPPORTED rate. The same mapping is applied to the teacher-forced solve on foreign answers in Table 3. Restricting the risk-control and test computations to agreed spans changes the test false-SUPPORTED rate from $2.3\%$ to $2.2\%$ and coverage by $0.4$ pp.

**Sensitivity to label noise.** The estimands are defined relative to adjudicated labels ($\kappa = 0.79$ between the two annotators before adjudication). To bound the effect of this noise we re-drew, in each of 1,000 replicates, the label of every claim on which the annotators disagreed from the two pre-adjudication scores with equal probability, and recomputed the test false-SUPPORTED rate at the selected threshold: it lies within $[1.9, 2.9]\%$ in every replicate, and re-running the selection procedure on the perturbed calibration labels certifies at $\vartheta = 0.03$ in $96\%$ of replicates. Under a harsher model in which $5\%$ of all labels are flipped at random, the rate lies within $[1.6, 3.6]\%$ and the procedure certifies in $88\%$ of replicates.

Table D.6: Test false-SUPPORTED rate at the selected threshold $\hat{\tau}_{\text{sup}} = 0.50$ by claim type and by benchmark family on the 1,031 audited test rows stamped SUPPORTED (prompt-clustered bootstrap 95% intervals span $\pm$1–5 pp for the per-type rows and $[1.5, 3.4]$ for the pooled rate). The per-type rates are diagnostics: the guarantee is marginal, and group-conditional control would require a per-type calibration set of the size that the sample-size analysis of Appendix D indicates (1,500 claims per group), which the pooled 2,000-claim set does not provide for any single type.

| Group | SUPPORTED rows | false-SUPPORTED | FSR (%) |
|---|---|---|---|
| Perception | 468 | 9 | 1.9 |
| OCR / numeric | 289 | 8 | 2.8 |
| Reasoning step | 165 | 4 | 2.4 |
| World knowledge | 109 | 3 | 2.7 |
| Reasoning / video benchmarks | 402 | 10 | 2.5 |
| OCR / chart / document benchmarks | 371 | 11 | 3.0 |
| Hallucination-sensitive benchmarks | 258 | 3 | 1.2 |
| All | 1,031 | 24 | 2.3 |

**Parser quality.** On the 400-prompt parser-eval set (1,200 annotated claims) two annotators independently produced atomic decompositions; against their union the atomizer's recall is $93.6\%$, precision $94.2\%$ (Appendix I). Claims the parser misses receive no row and are counted by the manual-atomization audit of Section 5.3, which is how parser error enters the headline numbers; it plays no role in Theorem 1.

**Blinded audit.** 520 prompts stratified over the fourteen benchmarks, paired at the prompt level, drawn with a fixed random seed before annotation and disjoint from all training and calibration splits. All outputs and checkpoints were frozen before annotation began. Annotators (not authors) saw randomised prompt–evidence–answer–claim packets with system names, certificates, verifier scores, and predicted verdicts hidden. The audit proceeded in two waves under one protocol, one annotator pool, and the same 520 prompts. Wave 1 covered the six systems of Table 2 (the backbone, on-policy grounder-DPO, on-policy GRPO, tool outputs in context with grounder-DPO, the grounder control, and the ledger); wave 2 covered the thirteen attribution generators of Table 3 (the answer-only LoRA, the four additive configurations, Controls A, B1, B2, and C, and the four portability rows), whose paired differences against the ledger reuse the ledger's wave-1 adjudications on the same prompts. For the four systems whose worst-case analysis is reported, excluded items were 137 unjudgeable claims ($2.2\%$ of 6,209 extracted; per system 34/35/36/32 for backbone / grounder-DPO / grounder control / ledger) and 53 atomization failures ($0.9\%$; 13/13/13/14). Exclusion is decided per claim, not per prompt, so the excluded counts differ across systems ($47/48/49/46$) although every system is scored on the same prompts; the worst-case analysis of Section 5.3 counts every excluded item against the ledger and for its comparator. Full-abstention answers (13/14/24/16) contribute no claims and enter answer coverage. Seed-level support for the attribution comparisons comes from the calibrated verifier: for B2$-$B1, C$-$ledger, and B1$-$ledger the three training seeds were matched by index and the paired differences bootstrapped over prompts within each seed, giving

$+0.9$ $[0.3, 1.5]$, $+0.7$ $[0.2, 1.2]$, and $+0.5$ $[-0.1, 1.1]$ pp; these three comparisons were fixed before the runs, and the first two remain positive after a Holm correction over the three. Paired differences use a prompt-clustered bootstrap with 10,000 resamples. The manual-atomization subsample (250 prompts) had annotators decompose answers themselves, blind to the atomizer output. The InternVL audit used a disjoint 200-prompt sample with the manual-atomization protocol and the same annotator pool. The twelve remaining distinct generators (the three frontier models, DoLa, VCD, OPERA, Woodpecker, LURE, HA-DPO, POVID, RLAIF-V, and tool outputs in context; the remaining controls trained on the same annotations share their answers with audited generators and inherit their hUCR) were audited on a 200-prompt shared subsample of the 520-prompt pool (≈580 judged claims per system, Wilson half-widths 2.3–2.9 pp); their hUCR values are reported in Table 1 or Figure A.3 and are used only as a grader-agreement check, never for a headline claim.

Table D.7: Reliability of the support score on the asserted rows (SUPPORTED or WEAK; NULL rows are decided by the null share and the contradiction flag before $\hat{s}$ is consulted) of the 1,500-claim calibration-test set, which is disjoint from the risk-control set used to select $\hat{\tau}_{\text{sup}}$. Each cell gives the share of asserted rows whose $\hat{s}$ falls in the bin and the mean adjudicated support of those rows. The ledger's scores concentrate in the top two bins (86% of asserted rows), which is why a 2.3% false-SUPPORTED rate at $\hat{\tau}_{\text{sup}} = 0.50$ coexists with calibrated scores; the $[0, .5)$ bin holds the 12 rows stamped WEAK by score (Appendix K). ECE is computed on ten equal-mass bins of the same population (2.4 / 2.3 / 1.4 for the three rows; bootstrap standard error 0.3). †: our frozen grounder attached to the backbone's answers.

| System / bin of $\hat{s}$ | $[0, .5)$ | $[.5, .8)$ | $[.8, .9)$ | $[.9, .95)$ | $[.95, .98)$ | $[.98, 1]$ |
|---|---|---|---|---|---|---|
| Qwen3-VL† | 5.2% / 0.36 | 8.6% / 0.70 | 13.0% / 0.84 | 18.4% / 0.91 | 26.0% / 0.95 | 28.8% / 0.98 |
| LoRA + grounder | 4.0% / 0.37 | 7.0% / 0.71 | 12.0% / 0.85 | 18.0% / 0.92 | 27.0% / 0.96 | 32.0% / 0.98 |
| LEDGER-VLM | 0.8% / 0.46 | 2.3% / 0.74 | 3.3% / 0.86 | 7.1% / 0.94 | 22.0% / 0.98 | 64.5% / 0.99 |

**Risk control under distribution shift.** The i.i.d. calibration condition of Theorem 1 fails by construction under the stress regimes, and the guarantee is expected to degrade. Re-scoring the audited slices of each regime at severity 0.75 with the in-distribution threshold $\hat{\tau}_{\text{sup}} = 0.50$ gives the test false-SUPPORTED rates of Table D.8; every value exceeds the in-distribution 2.3% and four of five exceed $\vartheta$. Re-running the selection procedure on 1,500 claims from the shifted regime restores the target at a coverage cost of 4–9 pp, which is the recalibration action of Appendix M. Sample size matters here: at $\vartheta = 0.03$ the fixed-sequence test needs enough SUPPORTED claims to reject $H_1$, and in bootstrap replicates of the modality-drop regime the procedure finds an admissible threshold in 91% of replicates with 1,500 claims but in only 12% with 400 (the remainder certify nothing and fall back to the deployment policy of Appendix M). We therefore recommend 1,500 claims per recalibration.

Table D.8: False-SUPPORTED rate and coverage under distribution shift (severity 0.75) with the in-distribution threshold, and after recalibrating $\hat{\tau}_{\text{sup}}$ on 1,500 shifted claims ($\vartheta = 0.03$).

| Regime | test FSR at $\tau = 0.50$ (%) | SUPPORTED coverage (%) | recalibrated $\hat{\tau}_{\text{sup}}$ | FSR / coverage after (%) |
|---|---|---|---|---|
| Conflict | 3.3 | 63.1 | 0.550 | 2.6 / 58.0 |
| Distractor | 2.8 | 64.8 | 0.525 | 2.4 / 60.4 |
| OCR noise | 3.6 | 61.2 | 0.575 | 2.7 / 55.9 |
| Modality drop | 4.4 | 52.6 | 0.625 | 2.8 / 43.8 |
| Temporal | 3.4 | 62.0 | 0.550 | 2.6 / 56.7 |

# E PROOFS

## E.1 SETTING

Fix a probability space on which the input $X$, the output $Y$, the evidence atomisation $E(X)$, the claim atomisation $C(Y) = \phi(Y)$, and the answer-unit decomposition $A(Y)$ are jointly defined, with $n, \ell, m \leq N$ almost surely. The evidence side carries the concrete marginal $u$ as an upper bound and a null row of free capacity (no marginal); feasibility is the partial-transport feasibility of Section 3.2, and the null source has a fixed positive cost $D_{\varnothing c} = \nu_0$, which is compared with the costs of assigning mass to concrete atoms (routing a whole column to the null source is optimal when every concrete candidate is dearer, Proposition 3). $\|\cdot\|_1$ is the entry-wise $\ell_1$ norm. The analysis takes the

composition guard $\epsilon_b = 0$; the implementation's $\epsilon_b = 10^{-6}$ changes the column sums of $\widehat{\Pi}^{\mathrm{EA}}$ by at most $\epsilon_b \ell$, which is below the reporting precision of every quantity here.

### E.2 The degree-overflow penalty

The soft-degree penalty of Section 3.2 uses $h(x) = x/(x + \varsigma_s)$ and reads

$$\Omega_s = \sum_{\Pi \in \{\Pi^{\mathrm{EC}}, \Pi^{\mathrm{CA}}, \Pi^{\mathrm{EA}}\}} \Big[ \sum_i \max\big(0, \textstyle\sum_j h(\Pi_{ij}/\beta_j) - K_r\big)^2 + \sum_j \max\big(0, \sum_i h(\Pi_{ij}/\beta_j) - K_c\big)^2 \Big], \tag{7}$$

where $\beta$ is the column budget of the plan ($b$ for $\Pi^{\mathrm{EC}}$, $q$ for $\Pi^{\mathrm{CA}}$ and $\Pi^{\mathrm{EA}}$), $\varsigma_s = 0.005$, $K_c = 8$, and $K_r = 2$. The scale matters: at $\varsigma_s = 0.05$ the column term could never exceed $K_c$ (the $K_{\mathrm{eff}} = 12$ candidates at share $1/12$ give $\sum h \approx 7.5$), whereas at $0.005$ a share of $0.05$ already counts $0.91$, so a column with three funding atoms at $0.3$ and nine candidates at $0.01$ has $\sum h \approx 8.9$ and is penalised while the same column with the nine at $10^{-3}$ has $4.5$ and is not. Entropic Sinkhorn keeps every supported entry positive and the support mask is stop-gradient, so a hard degree count would be constant within a forward pass and could not train; the soft degree is the differentiable surrogate. The hard count $d^{\mathrm{rep}} = |\{\Pi_{ij}/\beta_j \geq \varsigma\}|$ with $\varsigma = 10^{-3}$ is used only for reporting.

### E.3 Existence, uniqueness, and convergence of the semi-relaxed solve

**Lemma 1** (Fixed point of the semi-relaxed iteration)**.** *For fixed finite costs on a fixed support, the entropic problem* $\min_{\Pi \in \mathcal{U}} \langle D, \Pi \rangle + \varepsilon \sum_{ij} \Pi_{ij}(\log \Pi_{ij} - 1)$ *over the feasible set* $\mathcal{U}$ *of Section 3.2 (for* $\Pi^{\mathrm{EC}}$ *and* $\Pi^{\mathrm{EA}}$*: exact column budgets, concrete row sums bounded by* $u$*, free null row; for* $\Pi^{\mathrm{CA}}$*: exact row marginal* $b$ *and column marginal* $q$*) has a unique minimiser whenever* $\mathcal{U}$ *is non-empty, and the corresponding scaling iteration of Section 3.3 (semi-relaxed, respectively balanced) converges to it.*

*Proof. Existence and uniqueness.* $\mathcal{U}$ is a non-empty (for $\Pi^{\mathrm{EC}}$ and $\Pi^{\mathrm{EA}}$ the all-null plan is feasible; for $\Pi^{\mathrm{CA}}$ the span plan $\Pi^{\mathrm{CA}}_{jk} = q_k/n_k$ for $k \in \mathrm{span}(c_j)$, with $n_k$ the number of claims touching unit $k$, has row sums $\sum_{k \in \mathrm{span}(c_j)} q_k/n_k = b_j$ and column sums $\sum_{j:\, k \in \mathrm{span}(c_j)} q_k/n_k = q_k$ by the budget rule of Section 3.1, and its edges are always in the support, so the balanced problem is feasible; a proportional rule $b_j \propto \sum_{k \in \mathrm{span}(c_j)} q_k$ would not guarantee this, since two units of equal budget holding nine and eight claims would demand $9/17$ and $8/17$ of the mass on own-span edges), compact, convex polytope, and the objective is continuous and strictly convex on it because the negative entropy is strictly convex on the support, so a unique minimiser exists. *Convergence.* With $G = e^{-D/\varepsilon}$ on the support, the problem is $\min_\Pi \ \varepsilon\,\mathrm{KL}(\Pi \,\|\, G) + F_1(\Pi \mathbf{1}) + F_2(\Pi^\top \mathbf{1})$ up to an additive constant, where $F_1$ is the indicator of $\{a : \ a_i \leq u_i \text{ for every concrete row}\}$ (the null row is unconstrained) and $F_2$ is the indicator of $\{b\}$; both are proper, lower semicontinuous, convex functions. This is the setting of the scaling algorithms of Chizat et al. (2018): the iteration $a \leftarrow \mathrm{prox}^{\mathrm{KL}}_{F_1/\varepsilon}(Gv)\,/\,(Gv)$, $v \leftarrow \mathrm{prox}^{\mathrm{KL}}_{F_2/\varepsilon}(G^\top a)\,/\,(G^\top a)$ is exact block-coordinate ascent on the Fenchel dual, which is concave and continuously differentiable in each block because the conjugate of the entropy is smooth. The KL-proximal map of the indicator of $\{a \leq u\}$ is the componentwise map $\min(\cdot, u)$, which gives $a_i \leftarrow \min(1, u_i/(Gv)_i)$ for concrete rows and $a_\varnothing \equiv 1$ for the free row, and the KL-proximal map of the indicator of $\{b\}$ gives $v \leftarrow b/(G^\top a)$; these are the updates of Section 3.3. For $\Pi^{\mathrm{CA}}$ both $F_1$ and $F_2$ are indicators of equalities, the proximal maps give the classical balanced Sinkhorn updates $a \leftarrow b/(Gv)$, $v \leftarrow q/(G^\top a)$, and the same argument applies. A strictly positive feasible plan exists (spread an arbitrarily small mass over every supported concrete entry, which respects $u_i > 0$, and route the remainder of each column budget to the null row), so strong duality holds and the dual optimum is attained. Each block maximisation has a unique solution, so block-coordinate ascent on this concave, smooth, coercive dual converges to a dual maximiser (Tseng, 2001; Chizat et al., 2018), and the primal iterate $\Pi = \mathrm{diag}(a)\, G\, \mathrm{diag}(v)$ converges to the unique minimiser above. Note that the iteration is not an alternating KL projection in the sense of Bregman–Csiszár, since the row constraint is an inequality; the dual argument is what is needed. □

At the default $T = 15$ the plan differs from the converged solution ($T = 200$) by $1.2 \times 10^{-3}$ in entry-wise $\ell_1$ on average ($4.8 \times 10^{-3}$ at the 99th percentile); Proposition 1 is stated at the fixed point and holds to this precision for the unrolled solve.

### E.4 Proof of Proposition 1 and the measured binding fraction

The semi-relaxed iteration of Section 3.3 has the fixed-point form $\Pi^{\mathrm{EC}}_{ij} = a_i\, G_{ij}\, v_j$ on the support, with $G_{ij} = e^{-D_{ij}/\varepsilon}$, exact column scaling $v_j = b_j / \sum_{i \in S_j \cup \{\varnothing\}} a_i G_{ij}$, clipped concrete row scaling $a_i = \min(1, u_i / \sum_j G_{ij} v_j)$, and $a_\varnothing \equiv 1$. If no concrete row binds, $a_i = 1$ for every $i \le n$, and the column becomes $\Pi^{\mathrm{EC}}_{ij}/b_j = G_{ij} / \sum_{i' \in S_j \cup \{\varnothing\}} G_{i'j} = \mathrm{softmax}_{i' \in S_j \cup \{\varnothing\}}(-D_{i'j}/\varepsilon)_i$ with $D_{\varnothing j} = \nu_0$; nothing in this expression refers to another column, so the columns decouple. The same argument applies to $\Pi^{\mathrm{EA}}$ with answer units in place of claims. When some row binds, $a_i < 1$ multiplies that atom's kernel entries in every column it belongs to, which is the only channel by which claims interact. □

*Measured quantities.* At convergence on the evaluation set, 2.3% of concrete rows have $a_i < 1$ (median $a_i$ among binding rows 0.71), 6.1% of examples contain at least one binding row, and the mean entry-wise $\ell_1$ distance between a solved column of $\Pi^{\mathrm{EC}}$ and the softmax of Proposition 1 is 0.004 (0.31 on binding columns). Binding rows are concentrated on counting claims that share one object atom (41% of binding rows), on several claims competing for one OCR line (33%), and on duplicated objects (18%); Table G.20 reports UCR on these subsets for the ledger, the penalty-matched no-row-cap ledger, and Controls B1 and B2, with the split decided by the seed-17 ledger's fixed point over all evaluation prompts. Under demand-scaled capacities ($u_i$ rescaled by $\ell/n$ so that total supply is 1.2× the demand) rows bind on 38% of examples and the mean column distance to the softmax rises to 0.09; Table G.11 reports the corresponding variant.

### E.5 Proof of Theorem 1

Write $H_g : \mathrm{FSR}(\tau_g) > \vartheta$. Under the i.i.d. calibration condition of Theorem 1 the calibration claims are i.i.d. from $\mathcal{D}_{\mathrm{claim}}$ (sampling with replacement; a repeated prompt yields independent claim draws, not dependent claims), so conditional on $n_S(\tau_g)$ the number $k(\tau_g)$ of false-SUPPORTED calibration claims at $\tau_g$ is exactly $\mathrm{Bin}(n_S(\tau_g), \mathrm{FSR}(\tau_g))$, and $p_g = \Pr\big(\mathrm{Bin}(n_S(\tau_g), \vartheta) \le k(\tau_g)\big)$ is a valid $p$-value for $H_g$: whenever $H_g$ holds, $\Pr(p_g \le \delta) \le \delta$. The thresholds are tested in the fixed order $\tau_1 > \tau_2 > \dots$ and testing stops at the first non-rejection. Fixed-sequence testing controls the family-wise error rate at $\delta$ with no correction for the number of thresholds (Angelopoulos et al., 2021): the first true null in the sequence is falsely rejected with probability at most $\delta$, and no later hypothesis can be rejected unless every earlier one was, so on an event of probability at least $1 - \delta$ every rejected hypothesis is false. The selected $\hat{\tau}_{\mathrm{sup}}$ is a rejected index, hence $\mathrm{FSR}(\hat{\tau}_{\mathrm{sup}}) \le \vartheta$ on that event. No monotonicity of FSR in $\tau$ is used or needed, and no property of transport, of the null source, or of the score enters. If several claims of one answer were pooled, the binomial calculation would be invalid because they are dependent; that is why each draw keeps one claim and why draws are made with replacement, and why the test-set false-SUPPORTED rate, which uses every claim of 520 prompts, is reported with a prompt-clustered bootstrap interval rather than a binomial one. The guarantee is over the calibration draw and for deployment claims from the same $\mathcal{D}_{\mathrm{claim}}$; it says nothing about a stream whose claim distribution differs (Table D.8). □

**Why the joint rate is not bounded.** The joint event $\{\hat{v} = \text{SUPPORTED} \wedge s < \frac{1}{2}\}$ is contained in $\{s < \frac{1}{2}\}$, so its probability is at most the UCR; a bound expressed through parser and calibration residuals (about 0.2 with our residuals) is looser than that trivial bound and carries no information. The conditional rate has trivial bound 1 and is the quantity Theorem 1 controls.

**Coverage at controlled risk.** Table E.9 reports, for each system trained on the 6.2k annotations, the share of its own test claims stamped SUPPORTED when its threshold is chosen by the same procedure at $\vartheta = 0.03$, $\delta = 0.05$; the $2 \times 3$ cross of Table 3 repeats the measurement on shared answers. This, and not the guarantee itself, is where the ledger differs from the controls; most of the gap in Table E.9 reflects the ledger generator's lower unsupported rate, and the scorer's own margin on shared answers is the 1.8–2.1 pp of Table 3.

**What is not claimed.** The diagnostic $\Delta_{\mathrm{cyc}} + \Delta_{\mathrm{null}}$ is not asserted to bound total UCR. A qualitative routing assumption (each retained low-score claim contributes to null or cycle mass) cannot yield a nontrivial bound because a contribution can be arbitrarily small, and constants fit from data would be

Table E.9: Coverage at controlled risk ($\vartheta = 3\%$, $\delta = 0.05$, $n = 2{,}000$ calibration claims per system, each system on its own answers): threshold selected on the 40-point grid, calibration false-SUPPORTED count with the one-sided binomial $p$-value $\Pr\{\mathrm{Bin}(n_S, 0.03) \le k\}$ at the selected point (every earlier point of each sequence has a smaller $p$-value), test false-SUPPORTED rate on the audited seed-17 test claims, and the share of test claims stamped SUPPORTED (five-seed mean with SD in parentheses; the audited test labels exist for seed 17 only, so the false-SUPPORTED rate carries no seed SD).

| System (same 6.2k annotations) | $\hat{\tau}_{\mathrm{sup}}$ | calib. $k/n_S$ | $p$ at $\hat{\tau}_{\mathrm{sup}}$ | test FSR (%) | SUPPORTED coverage (%) |
|---|---|---|---|---|---|
| Verifier-only (LoRA answers) | 0.625 | 13/781 | 0.013 | 2.6 | 41.3 (1.12) |
| Post-hoc grounder (LoRA answers) | 0.575 | 16/912 | 0.013 | 2.7 | 47.6 (0.97) |
| Claim classifier (backbone answers) | 0.550 | 21/1,018 | 0.043 | 2.5 | 48.6 (0.88) |
| LEDGER-VLM | 0.500 | 22/1,296 | 0.002 | **2.3** | **68.2** (0.71) |

post hoc. The relation is tested empirically (Figure A.4c) and stated as an operational falsification test.

**Proposition 2** (Diagnostic decomposition)**.** *Each of $\Delta_{\mathrm{cyc}}, \Delta_{\mathrm{null}}, \Delta_{\mathrm{marg}}, \Delta_{\mathrm{sparse}}$ is the unique active term for some ledger configuration, so no term of Eq. (2) is algebraically redundant and each detects a distinct failure mode (route disagreement, unfunded content, marginal leakage, degree concentration).*

### E.6 PROOF OF PROPOSITION 2

Take $n = \ell = m = 2$, $u = b = q = (\frac{1}{2}, \frac{1}{2})$. *Pure cycle:* $\Pi^{\mathrm{EC}} = \Pi^{\mathrm{CA}} = \mathrm{diag}(\frac{1}{2}, \frac{1}{2})$, so $\widehat{\Pi}^{\mathrm{EA}} = \mathrm{diag}(\frac{1}{2}, \frac{1}{2})$; take $\Pi^{\mathrm{EA}}$ the swapped permutation $\frac{1}{2}\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$: marginals exact, no null mass, degrees $\le 1$, but $\Delta_{\mathrm{cyc}} = \|\Pi^{\mathrm{EA}} - \widehat{\Pi}^{\mathrm{EA}}\|_1 = 4 \times \frac{1}{2} = 2$. *Pure null:* route $c_1$ entirely to the free null row ($\Pi^{\mathrm{EC}}_{\varnothing,1} = \frac{1}{2}$) and $c_2$ to $e_1, e_2$ with $\frac{1}{4}$ each, so both column budgets are met and both evidence rows stay within their upper bounds $u = (\frac{1}{2}, \frac{1}{2})$; choose $\Pi^{\mathrm{CA}} = \mathrm{diag}(\frac{1}{2}, \frac{1}{2})$ and $\Pi^{\mathrm{EA}} = \widehat{\Pi}^{\mathrm{EA}}$: no constraint is violated, degrees $\le 2$, $\Delta_{\mathrm{cyc}} = 0$, and only $\Delta_{\mathrm{null}} > 0$. *Pure marginal:* $\Pi^{\mathrm{EC}} = \Pi^{\mathrm{EA}} = \widehat{\Pi}^{\mathrm{EA}}$ with row sums $\Pi^{\mathrm{EC}}\mathbf{1} = (0.6, 0.4)$ against the capacities $u = (\frac{1}{2}, \frac{1}{2})$, so the first row overflows by $0.1$ and only the capacity-overflow part of $\Gamma$, hence only $\Delta_{\mathrm{marg}}$, is positive. *Pure overflow:* column cap $K_c = 1$ (and $K_r = 1$) with $\Pi^{\mathrm{EC}} = \frac{1}{4}\mathbf{1}\mathbf{1}^\top$, exact marginals, and $\widehat{\Pi}^{\mathrm{EA}} = \Pi^{\mathrm{EA}}$: each soft degree exceeds the cap, so only $\Delta_{\mathrm{sparse}} > 0$. Each term is therefore the unique active term for some configuration, and removing it from Eq. (2) leaves that configuration unpenalised. □

**Proposition 3** (Null routing at the optimum)**.** *Fix a claim column $j$ with constant null cost $D_{\varnothing j} = \nu_0$. In the unregularised limit of Eq. (1) the optimal plan routes the whole budget $b_j$ to $\varnothing$ whenever $\min_{i \in S_j} D_{ij} > D_{\varnothing j}$; at finite $\varepsilon$ the null share of column $j$ is at least*

$$\left(1 + K_{\mathrm{eff}}\, e^{-(\min_{i \in S_j} D_{ij} - D_{\varnothing j})/\varepsilon}\right)^{-1}, \qquad K_{\mathrm{eff}} = K + K_{\mathrm{prior}}.$$

*An independent nearest-neighbour matcher without a null source must route the same column to $\arg\min_{i \in S_j} D_{ij}$ even when every candidate has a high matching cost. This is a standard property of dustbin transport (Sarlin et al., 2020; Chizat et al., 2018); it separates the ledger from matchers without a null source, not from the grounder control of Section 5, which has one.*

### E.7 PROOF OF PROPOSITION 3

Column $j$ of $\Pi^{\mathrm{EC}}$ must carry mass $b_j$ over $S_j \cup \{\varnothing\}$. The null row has no capacity constraint, and the concrete rows are constrained only from above ($\sum_j \Pi^{\mathrm{EC}}_{ij} \le u_i$), so the column subproblems are coupled only through those upper bounds. *Unregularised limit.* Because the null row is free, routing the whole of $b_j$ to $\varnothing$ is feasible for every column simultaneously, and a column pays $b_j D_{\varnothing j}$ for it; any concrete routing of column $j$ costs at least $b_j \min_{i \in S_j} D_{ij}$. Hence when $D_{\varnothing j} < \min_{i \in S_j} D_{ij}$ every optimal plan routes all of $b_j$ to $\varnothing$ regardless of the other columns (moving any mass to a concrete atom raises the cost of column $j$ and can only tighten a constraint elsewhere). *Finite $\varepsilon$.* The semi-relaxed Sinkhorn iteration scales columns exactly and clips the concrete row scalings so that their upper bounds hold; at a fixed point the concrete row scalings satisfy $a_i \le 1$ (a scaling of 1 for a slack row, strictly less for an active one) and the free null row has scaling exactly 1. The entries of

column $j$ are therefore $\Pi^{\mathrm{EC}}_{ij} \propto a_i e^{-D_{ij}/\varepsilon}$ with $a_i \le 1$ for $i \in S_j$ and $\Pi^{\mathrm{EC}}_{\varnothing j} \propto e^{-D_{\varnothing j}/\varepsilon}$, which gives the null share

$$\frac{\Pi^{\mathrm{EC}}_{\varnothing j}}{b_j} \;\ge\; \frac{e^{-D_{\varnothing j}/\varepsilon}}{e^{-D_{\varnothing j}/\varepsilon} + \sum_{i\in S_j} e^{-D_{ij}/\varepsilon}} \;\ge\; \Big(1 + K_{\mathrm{eff}}\, e^{-(\min_{i\in S_j} D_{ij} - D_{\varnothing j})/\varepsilon}\Big)^{-1},$$

using $|S_j| \le K_{\mathrm{eff}} = K + K_{\mathrm{prior}}$. The null mass enters $M_\varnothing$ and hence $\Delta_{\mathrm{null}}$ with weight $\lambda_\varnothing$; $\Delta_{\mathrm{null}}$ carries no gradient at all (Section 3.4): were it back-propagated, the first-order response would move $z^C_j$ towards the cheapest candidate, lowering $D_{ij}$ and the bound above without changing a token, which is the degenerate variant of Table G.11. Stopping the gradient prevents direct optimisation of the null-mass term; other losses can still change null allocations indirectly through the learned embeddings and costs, and the term reaches the generator only through the unlikelihood weights. The swap loss $\mathcal{L}_{\mathrm{cf}}$ compares the funding set before and after the decisive atom is replaced and penalises certificates whose funding set is unchanged. A matcher without a null source must place the whole budget of every column on concrete candidates, so it selects the nearest atom even when that atom is far; its cost grows with the distance, but it has no alternative to selecting a candidate. □

**Proposition 4** (Sparse complexity)**.** *With $K_{\mathrm{eff}} = K + K_{\mathrm{prior}}$ and $N = n + \ell + m$, the ledger runs in $O\big((TK_{\mathrm{eff}} + K^2_{\mathrm{eff}})N\big)$ time and $O(K_{\mathrm{eff}}N)$ memory, the $K^2_{\mathrm{eff}}$ term coming from composing two sparse plans by streaming, whereas dense attribution that materialises all evidence-to-answer interactions needs $\Omega(nm)$.*

### E.8 Proof of Proposition 4

Cost construction needs $O(dN)$ dot products. Per claim column the candidate set has at most $K + K_{\mathrm{prior}} = K_{\mathrm{eff}}$ entries, so each plan has at most $K_{\mathrm{eff}}N$ non-zeros and each Sinkhorn iteration costs $O(K_{\mathrm{eff}}N)$; $T$ iterations cost $O(TK_{\mathrm{eff}}N)$. Composing $\Pi^{\mathrm{EC}}\,\mathrm{diag}(b+\epsilon_b)^{-1}\Pi^{\mathrm{CA}}$ multiplies supports: each answer column of $\widehat{\Pi}^{\mathrm{EA}}$ has at most $K^2_{\mathrm{eff}}$ candidate entries, formed by streaming one column at a time and accumulating the $\ell_1$ difference to $\Pi^{\mathrm{EA}}$ over the union support, for $O(K^2_{\mathrm{eff}}N)$ time and $O(K_{\mathrm{eff}}N)$ memory. Any dense attribution naming all evidence-to-answer interactions must represent $nm$ quantities, hence $\Omega(nm)$. With $K = 8$, $K_{\mathrm{prior}} = 4$, $T = 15$, $N \le 512$, the solver accounts for $0.04$ of the measured $0.18$ overhead (profiled on the evaluation set: solver $0.04$, evidence tools and atomisation $0.09$, projector and verifier passes $0.03$, certificate serialisation $0.02$); the sensitivity rows of Table G.17 vary the solver share between $0.01$ ($K = 2$) and $0.11$ ($K = 16$) while the other components stay fixed. □

Propositions 2 and 4 are elementary; they are included so that the role of each gap term and the cost of the sparse solver are stated exactly rather than asserted.

## F Pseudocode as executed

## G Complete tables

Table G.10: **Fourteen-benchmark suite** ($\mathrm{Avg}_{14}$ is the exact mean of the fourteen cells with OCRBench divided by ten). Hall.Acc is HallusionBench question-pair accuracy (qAcc). Frozen systems are single greedy runs; trained systems are five-seed means (three for RLAIF-V), with per-cell seed dispersion in Appendix Table G.16. The claim classifier does not change the generator and inherits the backbone's task scores; the verifier and grounder controls share one answer-only LoRA generator. Decoding-time baselines change $\mathrm{Avg}_{14}$ by at most $0.2$ and are omitted; their HallusionBench and POPE scores are in Table 1.

| System | MMMU | M-Pro | MMB | MMStar | Doc | Text | Chart | AI2D | OCR | POPE | Hall.Acc | Vid | Ego | Ref | $\mathrm{Avg}_{14}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPT-4o | 70.3 | 53.6 | 82.1 | 69.5 | 91.1 | 77.4 | 86.7 | 84.6 | 736 | 89.8 | 55.0 | 77.2 | 72.4 | 87.4 | 76.48 |
| Llama-4-Maverick-17B-128E | 73.4 | 58.9 | 88.1 | 72.8 | 94.4 | 84.9 | 90.0 | 88.2 | 852 | 91.2 | 59.7 | 79.6 | 78.2 | 88.9 | 80.96 |
| InternVL3.5-241B-A28B | 77.7 | 61.5 | 91.0 | 76.0 | 96.9 | 86.4 | 91.2 | 90.3 | 918 | 93.3 | 63.1 | 82.4 | 81.5 | 91.4 | 83.89 |
| Qwen3-VL-235B-A22B (backbone) | 79.6 | 62.8 | 91.8 | 76.8 | 97.6 | 87.8 | 92.0 | 91.0 | 935 | 93.9 | 64.2 | 83.7 | 82.5 | 92.4 | 84.97 |
| + RLAIF-V | 80.0 | 63.1 | 92.0 | 77.1 | 97.6 | 87.8 | 92.1 | 91.0 | 935 | 96.1 | 69.5 | 83.8 | 82.6 | 92.5 | 85.62 |
| Answer-only LoRA (generator of the verifier / grounder controls) | 79.6 | 62.7 | 91.8 | 76.8 | 97.6 | 87.7 | 92.1 | 91.1 | 936 | 94.1 | 65.0 | 83.6 | 82.5 | 92.6 | 85.06 |
| Backbone + claim classifier (generator unchanged) | 79.6 | 62.8 | 91.8 | 76.8 | 97.6 | 87.8 | 92.0 | 91.0 | 935 | 93.9 | 64.2 | 83.7 | 82.5 | 92.4 | 84.97 |
| LEDGER-VLM (ours) | **81.2** | **64.3** | **93.0** | **78.4** | **97.8** | **89.0** | **93.5** | **92.1** | **942** | **97.4** | **74.6** | **84.7** | **83.3** | **93.7** | **86.94** |

This section holds the full ablation and replication tables referenced from Section 5, the per-metric dispersion behind Tables G.10–1, the per-seed values of the trained systems, and the scaling sweep.

**Algorithm 1** SPARSETRIOT: sparse tri-partite transport for one example

**Require:** $Z_E, Z_C, Z_A$; capacities $u$ (upper bounds), budgets $b, q$ (exact), null row free with constant cost $\nu_0$; $K$=8, $K_{\text{prior}}$=4, entropic $\varepsilon$=0.05, iterations $T$=15, soft-degree scale $\varsigma_s$=0.005, caps $K_c$=8, $K_r$=2

$D^{EC}, D^{CA}, D^{EA} \leftarrow$ costs of Section 3.3

For each target column: $\mathcal{N} \leftarrow \text{TopK}(-D, K) \cup \text{PROTECTEDPRIORS}(\leq K_{\text{prior}})$ ▷ selection is stop-gradient

$\Pi^{\text{EC}} \leftarrow \text{Sinkhorn}_{\text{semi}}(D^{EC}, u, b, \mathcal{N}, \varepsilon, T)$; $\Pi^{\text{EA}} \leftarrow \text{Sinkhorn}_{\text{semi}}(D^{EA}, u, q, \mathcal{N}, \varepsilon, T)$ ▷ semi-relaxed: exact columns, clipped concrete rows, free null row; unrolled; differentiable in $D$ on $\mathcal{N}$

$\Pi^{\text{CA}} \leftarrow \text{Sinkhorn}_{\text{bal}}(D^{CA}, b, q, \mathcal{N}^{CA}, \varepsilon, T)$ ▷ balanced: exact rows $b$ and columns $q$, no null row; $\mathcal{N}^{CA}$ always contains each claim's own span units, on which $\Pi^{\text{CA}}_{jk} = q_k/n_k$ is feasible by the budget rule

$\widehat{\Pi}^{\text{EA}} \leftarrow \Pi^{\text{EC}} \,\text{diag}(b + \epsilon_b)^{-1} \Pi^{\text{CA}}$ streamed per answer column over the union support

$M_\varnothing \leftarrow \sum_j \Pi^{\text{EC}}_{\varnothing j} + \sum_k \Pi^{\text{EA}}_{\varnothing k}$

$\Gamma \leftarrow \|\Pi^{\text{EC}\top}\mathbf{1} - b\|_1 + \|\Pi^{\text{CA}}\mathbf{1} - b\|_1 + \|\Pi^{\text{CA}\top}\mathbf{1} - q\|_1 + \|\Pi^{\text{EA}\top}\mathbf{1} - q\|_1 + \sum_{i\leq n}(\Pi^{\text{EC}}\mathbf{1} - u)^+_i + \sum_{i\leq n}(\Pi^{\text{EA}}\mathbf{1} - u)^+_i$ ▷ four equalities, two upper bounds

$\Omega_s \leftarrow \sum_\Pi \big[\sum_i \max(0, \tilde{d}^{\text{row}}_i(\Pi) - K_r)^2 + \sum_j \max(0, \tilde{d}^{\text{col}}_j(\Pi) - K_c)^2\big]$ with $\tilde{d} = \sum h(\Pi_{ij}/b_j)$, $h(x) = x/(x + \varsigma_s)$ ▷ soft degrees, Eq. (7)

$\Delta^{\nabla}_{\text{led}} \leftarrow \|\Pi^{\text{EA}} - \widehat{\Pi}^{\text{EA}}\|_1 + \lambda_g \Gamma + \lambda_s \Omega_s$; $\Delta_{\text{null}} \leftarrow \lambda_\varnothing \,\text{sg}(M_\varnothing)$ ▷ only $\Delta^{\nabla}_{\text{led}}$ carries gradient; $\Delta_{\text{led}} = \Delta^{\nabla}_{\text{led}} + \Delta_{\text{null}}$ is reported

**return** $(\Pi^{\text{EC}}, \Pi^{\text{CA}}, \Pi^{\text{EA}}, \widehat{\Pi}^{\text{EA}}, \Delta^{\nabla}_{\text{led}}, \Delta_{\text{null}})$

**Algorithm 2** TRAINLEDGERVLM: training with a frozen backbone

**Require:** $\mathcal{D}$=$\{(x, y^\star, E^\star)\}$; frozen backbone $f_\theta$; trainable $\Delta\theta$ (LoRA), $h_\psi$, $v_\omega$, fixed null cost $\nu_0$; weights $(\alpha, \beta, \gamma, \eta)$; onset $\tau_{\text{ul}}$=0.30

**for** minibatch $B$ **do**

  **for** $(x, y^\star, E^\star) \in B$ **do**

    $E \leftarrow \text{EVIDENCEATOMIZE}(x)$; $y \leftarrow f_{\theta+\Delta\theta}(x)$; $A \leftarrow \text{ANSWERUNITS}(y)$; $C \leftarrow \text{CLAIMATOMIZE}(y)$

    $(u, b, q) \leftarrow$ capacity and budget rules of Section 3.1: $q_k \propto \text{len}(a_k)$, $b_j = \sum_{k\in\text{span}(c_j)} q_k/n_k$; null row free, constant cost $\nu_0$

    $(\Pi, \widehat{\Pi}^{\text{EA}}, \Delta^{\nabla}_{\text{led}}, \Delta_{\text{null}}) \leftarrow \text{SPARSETRIOT}(h_\psi(E, C, A), u, b, q)$ ▷ $C$ from the text-side atomizer; $z^A$ from decoder states

    $\ell_{\text{ans}} \leftarrow -\log p_{\theta+\Delta\theta}(y^\star|x) - \sum_j \text{sg}\big((n^{\text{alloc}}_j - \tau_{\text{ul}})^+/(1 - \tau_{\text{ul}})\big) \sum_{t\in\text{tok}(c_j)} \log\big(1 - p_{\theta+\Delta\theta}(y_t|y_{<t}, x)\big)$ ▷ gold likelihood + ledger-weighted unlikelihood on the draft $y$; $\text{tok}(c_j)$ excludes tokens of claims whose null share is $\leq \tau_\varnothing$ under the current plans (no scorer, no risk-control threshold); $\ell_{\text{ver}} \leftarrow \text{VERIFIERLOSS}(v_\omega, \Pi^{\text{EC}}, C, E^\star)$

    for each swapped claim $j \in J_{\text{cf}}$ (half of the sequences; Eq. (6)): $i^\star_j \leftarrow \arg\max_i \Pi^{\text{EC}}_{ij}$ (stop-gradient); replace $z^E_{i^\star_j}$ by the embedding $z^E_{i'_j}$ of a same-modality atom of another example that does not support $c_j$; re-solve $(\tilde{\Pi}^{EC}, \tilde{\Pi}^{CA}, \tilde{\Pi}^{EA})$ with Algorithm 1 on the same $Z_C, Z_A$; $\ell_{\text{cf}} \leftarrow |J_{\text{cf}}|^{-1} \sum_{j\in J_{\text{cf}}} \big[\tilde{\Pi}^{EC}_{i'_j j}/b_j + \sum_k \tilde{\Pi}^{CA}_{jk} \tilde{\Pi}^{EA}_{i'_j k}/q_k\big]$ ▷ funding the swapped-in atom receives on the composed and on the direct route; gradients reach $z^E$, $z^C$, and $z^A$

    $\ell_{\text{cal}} \leftarrow \text{SUPPORTCALIBRATION}(\Pi^{\text{EC}}, E^\star)$

    $L(x) \leftarrow \ell_{\text{ans}} + \alpha\Delta^{\nabla}_{\text{led}} + \beta\ell_{\text{cf}} + \gamma\ell_{\text{ver}} + \eta\ell_{\text{cal}}$ ▷ $\ell_{\text{ver}}$: cross-entropy on $u_j$; $\ell_{\text{cal}}$: Brier loss on $\sigma(u_j)$

  **end for**

  update $(\Delta\theta, \psi, \omega)$ with AdamW on $|B|^{-1} \sum_x L(x)$ ▷ $\theta$ receives no gradient

**end for**

Two columns recur: UCR is the as-emitted rate of the generator, and aUCR is the selective rate over asserted claims at the listed coverage.

## H STRESS REGIMES

Tables H.22 and H.23 give the full five-regime sweep behind Figure A.1a: as-emitted strict UCR and task success for the backbone, on-policy grounder-DPO (the reference baseline; on-policy GRPO is within 0.2 pp of it on every unsupported-claim metric and is omitted for legibility), and the ledger at seven severities, with seed dispersion in the captions. Because UCR is computed over non-refused

**Algorithm 3** EMITCERTIFICATE: post-hoc certificate emission

**Require:** input $x$; trained modules; thresholds $(\tau_{\text{sup}}, \tau_{\varnothing}, \tau_{\text{cyc}}, \tau_{\text{contra}}) = (0.50, 0.50, 0.05, 0.25)$

$y \leftarrow f_{\theta+\Delta\theta}(x)$ (greedy; never edited); $E, A, C$ and plans as in Algorithm 1
**for** claim $c_j \in C$ **do**
  $F_j \leftarrow \{i \in S_j : \Pi^{\text{EC}}_{ij}/b_j \geq \zeta\}$; $\bar{c}_j \leftarrow \sum_{i\in F_j} \Pi^{\text{EC}}_{ij}/b_j$; $\bar{\delta}_j \leftarrow \sum_{i\in S_j\setminus F_j} \Pi^{\text{EC}}_{ij}/b_j$; $n^{\text{alloc}}_j \leftarrow \Pi^{\text{EC}}_{\varnothing j}/b_j$; renormalise so $\bar{c}_j + \bar{\delta}_j + n^{\text{alloc}}_j = 1$; $\text{flag}^{\varnothing}_j \leftarrow [n^{\text{alloc}}_j > \tau_{\varnothing}]$ ▷ null dominance, $\tau_{\varnothing} = 0.5$
  $r_j \leftarrow \big(\sum_k \Pi^{\text{CA}}_{jk} \| \Pi^{\text{EA}}_{:,k} - \widehat{\Pi}^{\text{EA}}_{:,k} \|_1\big) / \sum_k \Pi^{\text{CA}}_{jk}$ ▷ normalised residual
  $\text{flag}_j \leftarrow$ CONTRADICTION$(c_j, \{e_i : \Pi^{\text{EC}}_{ij}/b_j \geq \tau_{\text{contra}}\})$ ▷ a funding atom binds $c_j$'s (entity, attribute) to a value incompatible with $c_j$'s value; one atom suffices
  **if** $S_j = \emptyset$ **then** $n^{\text{alloc}}_j \leftarrow 1$; $\hat{s}_j \leftarrow 0$; $\hat{v}_j \leftarrow$ NULL ▷ no candidate atoms at all, e.g. modality drop
  **else** $u_j \leftarrow v_\omega(z^C_j, \{z^E_i\}_{i\in S_j}, \Pi^{\text{EC}}_{:,j})$; $\hat{s}_j \leftarrow g_{\text{iso}}(w^\top \varphi_j)$ with $\varphi_j = (u_j, \bar{c}_j, r_j, n^{\text{alloc}}_j, \max_i \Pi^{\text{EC}}_{ij}/b_j, \tilde{H}_j)$ of Eq. (4)
    **if** $\text{flag}^{\varnothing}_j \vee \text{flag}_j$ **then** $\hat{v}_j \leftarrow$ NULL
    **else if** $\hat{s}_j \geq \tau_{\text{sup}} \wedge r_j \leq \tau_{\text{cyc}}$ **then** $\hat{v}_j \leftarrow$ SUPPORTED
    **else** $\hat{v}_j \leftarrow$ WEAK
    **end if**
  **end if**
  append $(j, c_j, F_j, \bar{c}_j, \bar{\delta}_j, n^{\text{alloc}}_j, r_j, \hat{s}_j, \text{flag}_j, \hat{v}_j)$ ▷ the answer text is returned unchanged
**end for**
**return** $(y, \text{rows}, \Delta_{\text{led}}, M_{\varnothing})$

answers, the refusal rate is reported beside it; the ledger's refusal rate departs from the backbone's only where evidence is removed or contradicted.

**Perturbation recipes.** Every perturbation acts on the input that all systems receive, never on a module of one system, and its strength is a function of the severity $s \in \{0, 0.15, \ldots, 0.90\}$. *Conflict*: with probability $s$ a retrieved note that contradicts one adjudicated claim of the gold rationale (its value slot replaced by a plausible alternative drawn from the same benchmark) is prepended to the user turn as retrieved context for every system; for the ledger it also becomes a text atom priced by $\rho_{\text{mod}}$. *Distractor*: $\lceil 6s \rceil$ object crops from other images of the same benchmark are pasted into the image at random non-overlapping positions (each at most $8\%$ of the image area), so that the detector produces distractor atoms and every system sees the same composite image. *OCR noise*: text regions are degraded by Gaussian blur with $\sigma = 2.5s$ pixels and JPEG re-encoding at quality $100 - 70s$, which raises the OCR character error rate from $1.2\%$ to $23\%$ at $s = 0.9$. *Modality drop*: with probability $s$ the non-text modality of the prompt is removed (the image is replaced by a blank canvas of the same size; for video every frame), so that the tools produce no atoms and $S_j = \emptyset$. *Temporal*: for video prompts, frames are permuted within windows of $\lceil 8s \rceil$ consecutive frames and a fraction $s/2$ of frames is dropped. The same perturbed inputs are used for the backbone, grounder-DPO, and the ledger; Figure A.11 shows examples. The label-free diagnostics for the same 35 cells are plotted against J-ind UCR in Figure A.4c, and the row-level behaviour on real inputs is shown in Figure A.11.

# I ATOMIZER ERROR ANALYSIS

On the 400-prompt (1,200-claim) parser-eval set the evaluation atomizer's recall against gold decompositions is $93.6\%$, precision $94.2\%$, $F_1$ $93.9\%$; the ledger's trainable claim atomizer $h^C_\psi$, measured on the same set, reaches $92.9\%$ / $93.8\%$, and the two agree on $95.1\%$ of claim spans. Failure modes: *merge* (two claims as one, $3.1\%$), *split* (one claim as two, $2.6\%$), *drop* ($1.8\%$), *spurious* ($0.7\%$). The parser-disagreement diagnostic flags answers on which the two annotators differ by more than $0.10$ unmatched mass ($7.4\%$); Theorem 1 is reported with and without this filter and remains non-vacuous in both cases. The manual-atomization audit (Section 5.3) shows the headline gap does not depend on the atomizer on the audited subsample.

Table G.11: **Component ablations and controls** on Qwen3-VL-235B (full ledger: five seeds; every other retrained row: three seeds; dispersion in Appendix G). "Variant" rows execute alternative objectives or emitters; "Additive" rows start from answer-only LoRA and add one ledger term, taking any quantity they need from the ledger (the null shares that weight $\mathcal{L}_{\mathrm{ul}}$, the matched atoms of the InfoNCE alignment) from a frozen reference ledger (the seed-17 full model, whose projector and atomizer are not updated in these rows; the same reference provides the verdicts that define $\mathrm{tok}(c_j)$); "Control" rows are the attribution controls of Section 5.4: Control A trains the answer-only LoRA with the grounder-weighted unlikelihood term on the model's own drafts and nothing else (no $\Delta_{\mathrm{cyc}}$, $\Gamma$, $\Omega_s$, $\mathcal{L}_{\mathrm{cf}}$, $\mathcal{L}_{\mathrm{ver}}$, or $\mathcal{L}_{\mathrm{cal}}$; its certificate is the frozen grounder at matched coverage), Controls B1 and B2 keep every loss term except the row-degree part of $\Omega_s$, which presupposes a coupled solve, and replace the solver by the per-claim softmax with a dustbin of Proposition 1 (B2 also drops $\Pi^{\mathrm{EA}}$ and the composition), and Control C keeps every loss but replaces $\Delta_{\mathrm{cyc}}$ by an InfoNCE alignment of answer-unit states to their matched atoms; "Portability" rows add ledger terms to the on-policy grounder-DPO baseline; the grounder-weighted $\mathcal{L}_{\mathrm{ul}}$ of the third portability row is that of Control A. Gap is the gradient-carrying $\Delta^{\nabla}_{\mathrm{led}}$ (Eq. (2)) and is undefined ("–") for rows without a direct witness or without transport.

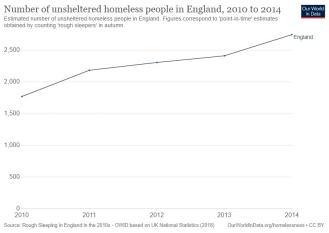
| Variant | Avg$_{14}$ ↑ | UCR↓ | aUCR↓ | Cert.$F_1$ ↑ | Cov.↑ | Gap$^{\nabla}$ ↓ | Lat. |
|---|---|---|---|---|---|---|---|
| Full ledger | **86.94** | **4.90** | **2.5** | **84.2** | **94.1** | **0.027** | 1.18× |
| − claim layer (direct $\Pi^{\mathrm{EA}}$ only; rows from $\Pi^{\mathrm{EA}}$ aggregated per claim span) | 84.23 | 9.50 | 7.2 | 72.1 | 86.0 | – | 1.12× |
| $-\Pi^{\mathrm{CA}}$ (composition replaced by span overlap) | 84.76 | 8.70 | 6.5 | 73.8 | 87.6 | 0.058 | 1.12× |
| $-\Pi^{\mathrm{EA}}$ (no direct witness; $r_j$ and $\Delta_{\mathrm{cyc}}$ dropped) | 83.94 | 10.10 | 7.9 | 70.4 | 84.7 | – | 1.10× |
| − null source (all columns fully funded; $\mathcal{L}_{\mathrm{ul}}$ weights vanish) | 84.58 | 12.30 | 12.3 | 61.3 | 100.0 | 0.048 | 1.09× |
| − counterfactual swaps | 85.43 | 7.80 | 5.6 | 76.2 | 88.8 | 0.043 | 1.16× |
| − ledger-weighted unlikelihood $\mathcal{L}_{\mathrm{ul}}$ | 85.72 | 7.30 | 4.6 | 79.8 | 91.4 | 0.036 | 1.18× |
| − soft-degree penalty | 85.84 | 6.90 | 5.0 | 77.9 | 90.1 | 0.039 | 1.19× |
| Variant: $\Gamma$ with four terms (no $\Pi^{\mathrm{EA}}$ marginals) | 86.87 | 5.00 | 2.7 | 83.9 | 94.0 | 0.031 | 1.18× |
| Variant: $\Delta_{\mathrm{null}}$ back-propagated into $z^C$ (degenerate) | 86.00 | 7.60 | 5.8 | 74.9 | 96.8 | 0.024 | 1.18× |
| Variant: draft weights synchronised every 50 steps | 86.96 | 4.80 | 2.5 | 84.3 | 94.2 | 0.027 | 1.18× |
| Gumbel-perturbed selection | 86.92 | 4.86 | 2.5 | 84.1 | 94.1 | 0.027 | 1.18× |
| Uniform evidence capacity $u \equiv 1$ | 86.63 | 5.40 | 3.1 | 83.4 | 93.8 | 0.029 | 1.18× |
| Learned per-atom capacity MLP | 86.96 | 4.90 | 2.5 | 84.3 | 94.1 | 0.027 | 1.21× |
| Variant: demand-scaled capacities (supply $\approx 1.2\times$ demand; rows bind on 38% of examples) | 86.76 | 5.10 | 2.7 | 83.1 | 93.9 | 0.031 | 1.19× |
| Variant: row cap $K_r = 1$ | 86.66 | 5.30 | 2.9 | 82.9 | 93.7 | 0.030 | 1.18× |
| Variant: row cap $K_r = 3$ | 86.95 | 4.90 | 2.5 | 84.0 | 94.1 | 0.028 | 1.18× |
| Variant: no row cap ($K_r = \infty$) | 86.54 | 5.60 | 3.2 | 81.7 | 93.5 | 0.033 | 1.19× |
| Additive: answer-only LoRA (no ledger terms, no certificate) | 85.06 | 10.80 | 10.8 | 73.2 | 100.0 | 0.061 | 1.04× |
| Additive: LoRA $+\mathcal{L}_{\mathrm{ul}}$ only (weights from the ledger's null shares) | 85.62 | 7.00 | 7.0 | 73.5 | 100.0 | 0.055 | 1.05× |
| Additive: LoRA $+\Delta_{\mathrm{cyc}}$ only via $\Pi^{\mathrm{EA}}$ (representation channel) | 85.70 | 8.80 | 7.7 | 74.1 | 93.0 | 0.049 | 1.15× |
| Additive: LoRA $+\mathcal{L}_{\mathrm{ul}} + \Delta_{\mathrm{cyc}}$ | 86.32 | 6.70 | 4.5 | 78.9 | 93.4 | 0.040 | 1.16× |
| Additive: LoRA $+\mathcal{L}_{\mathrm{cf}}$ only | 85.44 | 8.40 | 6.4 | 76.0 | 88.1 | 0.052 | 1.16× |
| Additive: LoRA $+\Delta_{\mathrm{led}}$ only | 85.56 | 8.10 | 6.0 | 75.8 | 88.5 | 0.044 | 1.16× |
| Additive: LoRA + InfoNCE alignment only (answer-unit states to matched atoms) | 85.64 | 8.30 | 7.3 | 73.9 | 92.8 | 0.051 | 1.15× |
| Additive: LoRA $+\mathcal{L}_{\mathrm{ul}}+$ InfoNCE alignment | 86.20 | 7.10 | 4.9 | 78.0 | 93.2 | 0.043 | 1.16× |
| Control A: drafts + grounder-weighted $\mathcal{L}_{\mathrm{ul}}$, no transport terms | 85.90 | 7.20 | 6.3 | 75.3 | 94.1 | – | 1.16× |
| Control B1: per-claim softmax + dustbin, $\Pi^{\mathrm{EA}}$ and $\Delta_{\mathrm{cyc}}$ kept | 86.59 | 5.40 | 3.1 | 81.9 | 93.8 | 0.034 | 1.15× |
| Control B2: per-claim softmax + dustbin, no composition, no $\Pi^{\mathrm{EA}}$ | 86.22 | 6.30 | 4.0 | 79.6 | 93.4 | – | 1.14× |
| Control C: full objective with InfoNCE alignment in place of $\Delta_{\mathrm{cyc}}$ | 86.50 | 5.60 | 3.1 | 82.0 | 93.7 | 0.035 | 1.17× |
| Portability: on-policy grounder-DPO $+\Delta_{\mathrm{cyc}}$ via $\Pi^{\mathrm{EA}}$ | 86.33 | 6.60 | 4.6 | 77.9 | 92.6 | 0.046 | 1.14× |
| Portability: on-policy grounder-DPO $+\Delta_{\mathrm{cyc}} + \mathcal{L}_{\mathrm{cf}}$ | 86.54 | 6.10 | 4.1 | 79.4 | 93.0 | 0.041 | 1.15× |
| Portability: on-policy grounder-DPO + grounder-weighted $\mathcal{L}_{\mathrm{ul}}$ | 86.46 | 6.30 | 4.4 | 78.1 | 92.9 | – | 1.14× |
| Portability: on-policy grounder-DPO $+\Delta_{\mathrm{cyc}} + \mathcal{L}_{\mathrm{cf}} + \mathcal{L}_{\mathrm{ul}}$ | 86.72 | 5.60 | 3.5 | 80.6 | 93.4 | 0.036 | 1.15× |
| Answer-only LoRA + post-hoc grounder (same 6.2k) | 85.06 | 10.80 | 7.6 | 74.0 | 86.4 | 0.062 | 1.23× |

Table G.12: **Architecture-distinct replication** on InternVL3.5-38B (`OpenGVLab/InternVL3_5-38B`; three seeds; seed SD in parentheses; strict UCR under the J-ind grader for all systems; device-hours are active H100-hours per seed). The transfer ablation keeps the Qwen-trained claim atomizer (token input) and the verifier head behind a re-fit linear input map, and retrains only the projector and adapters.

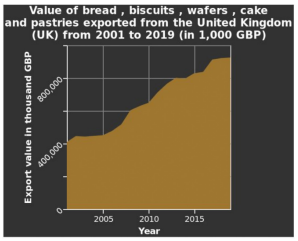
| System | UCR↓ | Hall.Acc↑ | POPE $F_1$ ↑ | Cert.$F_1$ ↑ | Claim cov.↑ | Lat. |
|---|---|---|---|---|---|---|
| InternVL3.5-38B, matched answer-only LoRA | 12.60 (0.24) | 60.1 (0.48) | 92.6 (0.19) | 68.4 (0.57)$^{\dagger}$ | 91.0 (0.42)$^{\dagger}$ | 1.00× |
| + Ledger, transfer ablation (456 device-h / seed) | 6.40 (0.29) | 67.8 (0.66) | 95.2 (0.28) | 78.6 (0.52) | 93.1 (0.39) | 1.17× (0.02) |
| + Ledger, full retraining (1,201 device-h / seed) | **5.90** (0.21) | **69.3** (0.61) | **96.0** (0.26) | **80.2** (0.44) | **93.8** (0.31) | 1.17× (0.02) |

# J CERTIFICATE GALLERY ON REAL BENCHMARK INPUTS

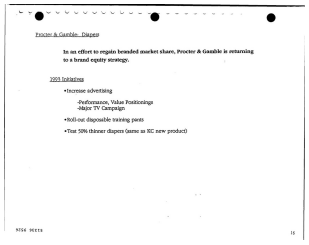
Figure J.12 shows certificates emitted for inputs drawn from ChartQA, HallusionBench, DocVQA, TextVQA, POPE, and AI2D. Each row lists the claim, the funding evidence identifiers (atoms at share $\geq \zeta = 0.10$), the concentrated share $\bar{c}_j$, the diffuse share $\bar{\delta}_j$, the null allocation $n_j^{\mathrm{alloc}}$ (the three sum to one), the residual $r_j$, the calibrated support $\hat{s}_j$, and the verdict; every value is read from the emitted certificate JSON, and the three inputs that also appear in Figure A.9 show the same numbers there. Figures A.8–A.11 show, in order, the counterfactual masking test on a real photograph, the transport plans behind three of these certificates, six further qualitative examples with their original benchmark questions, and the row-level behaviour under the stress regimes of Section 5.5.




| Claim | Funding atoms | $\bar{c}_j$ | $\bar{\delta}_j$ | $n_j^{\text{alloc}}$ | $r_j$ | $\hat{s}_j$ | Verdict |
|---|---|---|---|---|---|---|---|
| "The count of unsheltered people in England rose every year from 2010 to 2014" | E_series, E_axis_x, E_title | 0.91 | 0.06 | 0.03 | 0.019 | 0.88 | SUPPORTED |
| "The 2014 value exceeds 2,700" | E_series, E_axis_y | 0.86 | 0.10 | 0.04 | 0.024 | 0.81 | SUPPORTED |
| "The increase accelerated after 2015" | E_axis_x (0.19), E_series (0.10) | 0.29 | 0.06 | 0.65 | 0.014 | 0.06 | NULL |




| Claim | Funding atoms | $\bar{c}_j$ | $\bar{\delta}_j$ | $n_j^{\text{alloc}}$ | $r_j$ | $\hat{s}_j$ | Verdict |
|---|---|---|---|---|---|---|---|
| "UK exports of bread, biscuits, and pastries grew from 2001 to 2019" | E_chart_line, E_axis_x | 0.86 | 0.06 | 0.08 | 0.021 | 0.84 | SUPPORTED |
| "The value is reported in thousand GBP" | E_axis_y_label | 0.79 | 0.09 | 0.12 | 0.016 | 0.77 | SUPPORTED |
| "2010 was the year with the highest export value" | E_point_2010 (0.30); E_point_2019 (0.28) binds the maximum to 2019: contradiction flag | 0.58 | 0.00 | 0.42 | 0.058 | 0.37 | NULL |



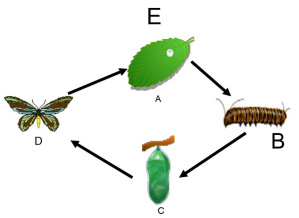

| Claim | Funding atoms | $\bar{c}_j$ | $\bar{\delta}_j$ | $n_j^{\text{alloc}}$ | $r_j$ | $\hat{s}_j$ | Verdict |
|---|---|---|---|---|---|---|---|
| "The memo describes a strategy to regain branded market share" | E_paragraph, E_heading | 0.90 | 0.07 | 0.03 | 0.022 | 0.89 | SUPPORTED |
| "A field-training campaign is one listed initiative" | E_initiatives | 0.84 | 0.11 | 0.05 | 0.027 | 0.82 | SUPPORTED |
| "The memo is signed by the division president" | ∅ (no atom at share $\geq \zeta$) | 0.00 | 0.12 | 0.88 | 0.012 | 0.05 | NULL |




| Claim | Funding atoms | $\bar{c}_j$ | $\bar{\delta}_j$ | $n_j^{\text{alloc}}$ | $r_j$ | $\hat{s}_j$ | Verdict |
|---|---|---|---|---|---|---|---|
| "The sign reads 'Denny's' " | E_ocr_sign_01 | 0.92 | 0.07 | 0.01 | 0.015 | 0.93 | SUPPORTED |
| "The letters are yellow on a red background" | E_ocr_sign_01, E_region_sign | 0.83 | 0.05 | 0.12 | 0.020 | 0.80 | SUPPORTED |
| "The restaurant is open 24 hours" | ∅ (no atom at share $\geq \zeta$) | 0.00 | 0.04 | 0.96 | 0.011 | 0.04 | NULL |



| Claim | Funding atoms | $\bar{c}_j$ | $\bar{\delta}_j$ | $n_j^{\text{alloc}}$ | $r_j$ | $\hat{s}_j$ | Verdict |
|---|---|---|---|---|---|---|---|
| "A sandwich sits on the plate" | E_box_sandwich, E_box_plate | 0.92 | 0.02 | 0.06 | 0.017 | 0.93 | SUPPORTED |
| "A side salad accompanies the sandwich" | E_box_salad, E_box_plate | 0.89 | 0.06 | 0.05 | 0.023 | 0.86 | SUPPORTED |
| "A bowl is present in the image" | ∅ (no atom at share $\geq \zeta$) | 0.00 | 0.07 | 0.93 | 0.013 | 0.05 | NULL |




| Claim | Funding atoms | $\bar{c}_j$ | $\bar{\delta}_j$ | $n_j^{\text{alloc}}$ | $r_j$ | $\hat{s}_j$ | Verdict |
|---|---|---|---|---|---|---|---|
| "Label D denotes the adult butterfly stage" | E_diagram_D, E_arrow_CD | 0.85 | 0.05 | 0.10 | 0.020 | 0.86 | SUPPORTED |
| "The cycle proceeds egg → larva → pupa → adult" | E_arrow_EB, E_arrow_BC, E_arrow_CD | 0.74 | 0.10 | 0.16 | 0.031 | 0.72 | SUPPORTED |
| "The diagram shows five life stages" | E_diagram_D | 0.38 | 0.21 | 0.41 | 0.046 | 0.44 | WEAK |

Figure J.12: **Certificate gallery on real inputs** (ChartQA, HallusionBench, DocVQA, TextVQA, POPE, AI2D). Null verdicts localise the unsupported assertion (a year absent from the chart, a signature absent from the scan, a bowl absent from the photo) without refusing the whole answer; the HallusionBench row shows the contradiction flag: a single funding atom above $\tau_{\text{contra}}$ binds the maximum to a different year, so the row is NULL although its null share (0.42) is below $\tau_{\varnothing}$; the AI2D WEAK row has $n_j^{\text{alloc}} = 0.41 < \tau_{\varnothing}$ and $\hat{s}_j < \hat{\tau}_{\text{sup}}$ with $r_j \leq \tau_{\text{cyc}}$, so it is WEAK by score, the rarer of the two routes to WEAK (Appendix K).

Table G.13: Reliability metrics, part A: mean (SD). For trained systems (ours, the four matched-data controls, VCD's noise seeds: five seeds $\{17, 29, 41, 53, 67\}$) the SD is the sample SD over training seeds (LURE, HA-DPO, POVID, RLAIF-V, the on-policy baselines: three seeds $\{17, 29, 41\}$). Frozen or deterministic systems (marked †: API models, the unmodified backbone, DoLa, OPERA, Woodpecker, tool outputs in context) are single runs decoded greedily; their SD comes from a 10,000-resample prompt-level bootstrap of the evaluation set and describes sampling uncertainty, not run-to-run variance. Part B (Table G.14) covers the remaining metrics; per-seed values for the two headline systems are in Table G.15.

| System | $\mathrm{UCR}_{\mathrm{ver}}$ | $\mathrm{UCR}_{\mathrm{J\text{-}ind}}$ | Cert.$F_1$ | ECE | Coverage |
|---|---|---|---|---|---|
| GPT-4o (`gpt-4o-2024-11-20`) | 12.40 (0.16)† | 12.90 (0.13)† | 63.10 (0.72)† | 2.80 (0.19)† | 78.40 (0.45)† |
| Llama-4-Maverick-17B-128E | 13.10 (0.17)† | 13.60 (0.16)† | 66.80 (0.61)† | 2.90 (0.19)† | 81.00 (0.41)† |
| InternVL3.5-241B-A28B | 11.60 (0.13)† | 12.20 (0.19)† | 69.80 (0.66)† | 2.50 (0.16)† | 83.40 (0.42)† |
| Qwen3-VL-235B-A22B (backbone) | 11.40 (0.13)† | 11.80 (0.13)† | 70.60 (0.68)† | 2.40 (0.17)† | 84.20 (0.50)† |
| + DoLa | 10.60 (0.14)† | 10.90 (0.18)† | 71.00 (0.60)† | 2.40 (0.20)† | 84.50 (0.43)† |
| + VCD | 10.10 (0.21) | 10.82 (0.20) | 71.40 (0.33) | 2.33 (0.18) | 84.82 (0.69) |
| + OPERA | 9.90 (0.15)† | 10.00 (0.17)† | 71.60 (0.56)† | 2.30 (0.18)† | 84.60 (0.45)† |
| + Woodpecker | 9.00 (0.18)† | 9.60 (0.13)† | 71.90 (0.71)† | 2.30 (0.19)† | 84.90 (0.33)† |
| + LURE | 9.49 (0.25) | 10.14 (0.35) | 72.20 (1.00) | 2.25 (0.11) | 85.10 (0.44) |
| + HA-DPO | 9.11 (0.26) | 9.30 (0.39) | 73.93 (0.59) | 2.29 (0.07) | 86.89 (0.53) |
| + POVID | 8.72 (0.41) | 9.25 (0.28) | 74.19 (0.55) | 2.22 (0.07) | 87.19 (0.41) |
| + RLAIF-V | 7.89 (0.30) | 8.24 (0.27) | 75.10 (0.51) | 2.08 (0.15) | 87.77 (0.73) |
| + on-policy grounder-DPO (own drafts, matched compute) | 7.63 (0.20) | 7.77 (0.34) | 75.58 (0.94) | 2.10 (0.10) | 88.15 (0.36) |
| + on-policy GRPO (grounder claim reward, matched compute) | 7.38 (0.26) | 7.57 (0.35) | 75.92 (0.45) | 2.07 (0.06) | 88.42 (0.52) |
| + tool outputs in context (OCR, detector, chart text given to the decoder) | 9.60 (0.18)† | 9.90 (0.20)† | 72.00 (0.51)† | 2.30 (0.20)† | 85.30 (0.33)† |
| + tool outputs in context + on-policy grounder-DPO | 7.09 (0.34) | 7.26 (0.36) | 75.83 (0.84) | 2.07 (0.15) | 88.32 (0.31) |
| Answer-only LoRA + verifier head (same 6.2k) | 10.84 (0.37) | 11.04 (0.37) | 73.23 (1.05) | 2.27 (0.07) | 85.04 (0.50) |
| Answer-only LoRA + post-hoc grounder (same 6.2k) | 10.84 (0.37) | 11.04 (0.37) | 74.04 (0.58) | 2.30 (0.18) | 86.41 (0.37) |
| same, abstention head @ Cov = 94.1 | 10.84 (0.37) | 11.04 (0.37) | 76.14 (0.52) | 2.06 (0.11) | 94.14 (0.53) |
| Backbone + claim classifier @ 94.1 (same 6.2k) | 11.40 (0.13)† | 11.80 (0.13)† | 76.37 (0.48) | 2.24 (0.12) | 94.13 (0.40) |
| LEDGER-VLM (ours) | **4.91 (0.19)** | **5.31 (0.30)** | **84.23 (0.25)** | **1.36 (0.08)** | **94.08 (0.22)** |

Table G.14: Reliability metrics, part B: same conventions as Table G.13. Latency is wall-clock relative to the unmodified backbone served locally; its SD is over five timing runs for every system; the API model has no comparable latency.

| System | Hall.Acc | POPE $F_1$ | ME-AUC | Latency |
|---|---|---|---|---|
| GPT-4o (`gpt-4o-2024-11-20`) | 55.00 (1.27)† | 89.80 (0.49)† | 75.20 (0.79)† | API |
| Llama-4-Maverick-17B-128E | 59.70 (1.34)† | 91.20 (0.39)† | 77.00 (0.70)† | 0.96× (0.01)† |
| InternVL3.5-241B-A28B | 63.10 (1.23)† | 93.30 (0.35)† | 79.20 (0.67)† | 1.02× (0.01)† |
| Qwen3-VL-235B-A22B (backbone) | 64.20 (1.75)† | 93.90 (0.38)† | 79.80 (0.78)† | 1.00× (0.01)† |
| + DoLa | 65.10 (1.32)† | 94.40 (0.35)† | 80.10 (0.68)† | 1.09× (0.01)† |
| + VCD | 66.00 (0.52) | 94.86 (0.14) | 80.63 (0.55) | 1.92× (0.03) |
| + OPERA | 66.40 (1.67)† | 95.10 (0.38)† | 80.90 (0.77)† | 1.41× (0.02)† |
| + Woodpecker | 66.20 (1.23)† | 95.00 (0.47)† | 81.20 (0.79)† | 2.63× (0.02)† |
| + LURE | 66.89 (0.52) | 95.26 (0.19) | 81.44 (0.51) | 1.36× (0.01) |
| + HA-DPO | 67.05 (0.83) | 95.41 (0.28) | 81.89 (0.33) | 1.00× (0.03) |
| + POVID | 67.93 (0.53) | 95.62 (0.26) | 82.42 (0.35) | 1.00× (0.01) |
| + RLAIF-V | 69.54 (0.41) | 96.12 (0.21) | 83.62 (1.06) | 1.00× (0.02) |
| + on-policy grounder-DPO (own drafts, matched compute) | 70.37 (0.74) | 96.26 (0.16) | 83.99 (0.48) | 1.00× (0.01) |
| + on-policy GRPO (grounder claim reward, matched compute) | 70.94 (0.39) | 96.44 (0.17) | 84.31 (1.10) | 1.00× (0.02) |
| + tool outputs in context (OCR, detector, chart text given to the decoder) | 66.80 (1.52)† | 95.20 (0.35)† | 81.00 (0.81)† | 1.12× (0.01)† |
| + tool outputs in context + on-policy grounder-DPO | 71.17 (0.59) | 96.54 (0.15) | 84.41 (0.46) | 1.12× (0.02) |
| Answer-only LoRA + verifier head (same 6.2k) | 64.96 (0.50) | 94.07 (0.18) | 80.43 (0.47) | 1.04× (0.02) |
| Answer-only LoRA + post-hoc grounder (same 6.2k) | 64.96 (0.50) | 94.07 (0.18) | 81.67 (0.66) | 1.23× (0.02) |
| same, abstention head @ Cov = 94.1 | 64.96 (0.50) | 94.07 (0.18) | 82.98 (0.39) | 1.27× (0.02) |
| Backbone + claim classifier @ 94.1 (same 6.2k) | 64.20 (1.75)† | 93.90 (0.38)† | 82.07 (0.78) | 1.06× (0.02) |
| LEDGER-VLM (ours) | **74.63 (0.44)** | **97.36 (0.12)** | **91.58 (0.48)** | 1.18× (0.02) |

# K VERDICT CONFUSION, ERROR ANALYSIS, AND THE TASK-SCORE MECHANISM

**Error analysis of the 35 asserted unsupported claims.** The 24 false-SUPPORTED and 11 false-WEAK rows of Table K.24 decompose as follows: 15 *value errors* (a correct entity and attribute with a wrong number, date, or count; 11 of them numeric readings from charts or documents whose funding atom was the correct series or line, so the funding set is right and the value is not, which is exactly what the strict criterion catches and a similarity-based cost cannot see), 7 *wrong attributes* (colour, material, or state of a correctly identified object), 6 *invented relations* between two correctly

Table G.15: Per-run values behind Table G.13 for the trained systems audited on the full 520-prompt sample (seeds 17/29/41/53/67; every value is one complete fine-tuning run followed by the same evaluation). The verifier and grounder controls share one answer-only LoRA generator per seed, so their generator metrics (UCR, Hall.Acc, POPE) are listed once; only their certificate metrics differ. The three on-policy baselines were trained with seeds 17/29/41 only. The last two columns recompute the mean and sample SD from the listed entries. The frozen backbone (and the claim classifier, which reuses its answers) is a single greedy run and has no per-seed row. Seed 17, the audited seed, is the ledger's weakest seed and the answer-only LoRA generator's strongest; the human audit uses seed 17 for every system.

| Metric | seed 17 | seed 29 | seed 41 | seed 53 | seed 67 | mean | SD |
|---|---|---|---|---|---|---|---|
| LEDGER-VLM | | | | | | | |
| $\text{UCR}_{\text{ver}}$ (%) | 5.23 | 4.74 | 4.81 | 4.96 | 4.83 | 4.91 | 0.19 |
| Cert.$F_1$ | 83.96 | 84.20 | 84.36 | 84.60 | 84.05 | 84.23 | 0.25 |
| Coverage (%) | 94.22 | 93.71 | 94.06 | 94.14 | 94.25 | 94.08 | 0.22 |
| ECE (%) | 1.26 | 1.33 | 1.36 | 1.47 | 1.37 | 1.36 | 0.08 |
| Hall.Acc | 74.19 | 74.44 | 75.19 | 74.34 | 74.99 | 74.63 | 0.44 |
| POPE $F_1$ | 97.45 | 97.26 | 97.53 | 97.26 | 97.30 | 97.36 | 0.12 |
| *Answer-only LoRA generator (shared by the verifier and grounder controls): generator metrics* | | | | | | | |
| $\text{UCR}_{\text{ver}}$ (%) | 10.41 | 11.06 | 10.61 | 10.77 | 11.35 | 10.84 | 0.37 |
| Hall.Acc | 65.30 | 65.25 | 65.08 | 65.07 | 64.08 | 64.96 | 0.50 |
| POPE $F_1$ | 93.91 | 94.00 | 93.93 | 94.35 | 94.14 | 94.07 | 0.18 |
| *+ grounder @ Cov 94.1: certificate metrics* | | | | | | | |
| Cert.$F_1$ | 75.87 | 77.04 | 75.81 | 75.85 | 76.14 | 76.14 | 0.52 |
| Coverage (%) | 94.13 | 94.81 | 94.26 | 94.16 | 93.33 | 94.14 | 0.53 |
| ECE (%) | 1.88 | 2.05 | 2.18 | 2.09 | 2.12 | 2.06 | 0.11 |
| *+ verifier head: certificate metrics* | | | | | | | |
| Cert.$F_1$ | 74.35 | 73.83 | 72.02 | 72.21 | 73.76 | 73.23 | 1.05 |
| Coverage (%) | 85.26 | 85.56 | 84.94 | 85.19 | 84.24 | 85.04 | 0.50 |
| ECE (%) | 2.29 | 2.37 | 2.17 | 2.25 | 2.25 | 2.27 | 0.07 |
| *On-policy grounder-DPO (three seeds): generator metrics* | | | | | | | |
| $\text{UCR}_{\text{ver}}$ (%) | 7.74 | 7.75 | 7.40 | – | – | 7.63 | 0.20 |
| Hall.Acc | 70.54 | 69.56 | 71.02 | – | – | 70.37 | 0.74 |
| POPE $F_1$ | 96.21 | 96.43 | 96.13 | – | – | 96.26 | 0.16 |
| *On-policy GRPO (three seeds): generator metrics* | | | | | | | |
| $\text{UCR}_{\text{ver}}$ (%) | 7.08 | 7.54 | 7.52 | – | – | 7.38 | 0.26 |
| Hall.Acc | 71.11 | 71.21 | 70.49 | – | – | 70.94 | 0.39 |
| POPE $F_1$ | 96.62 | 96.28 | 96.42 | – | – | 96.44 | 0.17 |
| *Tool outputs in context + grounder-DPO (three seeds): generator metrics* | | | | | | | |
| $\text{UCR}_{\text{ver}}$ (%) | 6.76 | 7.09 | 7.43 | – | – | 7.09 | 0.34 |
| Hall.Acc | 71.77 | 70.59 | 71.14 | – | – | 71.17 | 0.59 |
| POPE $F_1$ | 96.57 | 96.68 | 96.38 | – | – | 96.54 | 0.15 |

Table G.16: Per-cell seed standard deviations for the trained rows of Table G.10 (same units as the means; OCRBench in native points; $\text{Avg}_{14}$ SD over seed-level averages). Frozen systems and the claim-classifier control (which does not change the generator) are single greedy runs and have no seed dispersion; their prompt-bootstrap SDs are in Table G.13.

| System | MMMU | M-Pro | MMB | MMStar | Doc | Text | Chart | AI2D | OCR | POPE | Hall.Acc | Vid | Ego | Ref | $\text{Avg}_{14}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LEDGER-VLM (ours) | 0.14 | 0.22 | 0.16 | 0.19 | 0.28 | 0.21 | 0.12 | 0.22 | 2 | 0.12 | 0.44 | 0.24 | 0.12 | 0.31 | 0.10 |
| + RLAIF-V (3 seeds) | 0.22 | 0.34 | 0.24 | 0.25 | 0.36 | 0.36 | 0.35 | 0.41 | 4 | 0.21 | 0.41 | 0.23 | 0.29 | 0.48 | 0.15 |
| Answer-only LoRA (same 6.2k) | 0.37 | 0.39 | 0.26 | 0.39 | 0.18 | 0.18 | 0.43 | 0.40 | 3 | 0.18 | 0.50 | 0.40 | 0.32 | 0.28 | 0.16 |

detected objects ("the man holds the umbrella"), 4 *world-knowledge errors* that no atom could refute, and 3 *atomization artefacts* in which a merged span carried one correct and one incorrect proposition. Value errors are the dominant residual failure: relevance is not entailment, and the ledger's costs measure relevance. The contradiction flag is the only component that reads values, and it catches 19 of the 34 adjudicated value errors on the audit; extending it from exact numeric mismatch to tolerance-aware comparison and to categorical slots beyond the current lexicon is the most direct route to a lower asserted-claim error rate.

**What the verdicts do and do not tell an auditor.** SUPPORTED rows are wrong 2.3% of the time and WEAK rows 2.8% (11/392), so WEAK is a low-confidence assertion rather than a second risk level, and the deployment policy treats the two alike; the useful distinction is between asserted rows and NULL rows, and even NULL is a pointer (precision 39%, recall 50%) whose value is measured

Table G.17: Hyper-parameter and threshold sensitivity ($K$=8, $T$=15, thresholds $(\hat{\tau}_{\text{sup}}, \tau_{\varnothing}, \tau_{\text{cyc}}) = (0.50, 0.50, 0.05)$ are defaults; $K$ and $T$ rows are retrained with three seeds each and are costed in Table C.4; seed SD $\leq 0.17$ for $\text{Avg}_{14}$ and $\leq 0.4$ pp for UCR and aUCR). Latency follows the measured overhead breakdown of Appendix E (solver share 0.04 of 0.18 at default, 0.01 at $K$=2, 0.11 at $K$=16). $\tau_{\varnothing}$ is applied at emission only: it moves rows between NULL and the asserted verdicts and therefore changes coverage, aUCR, and Cert.$F_1$ while leaving $\text{Avg}_{14}$ and the as-emitted UCR untouched, because the certificate never edits the answer. $\hat{\tau}_{\text{sup}}$ and $\tau_{\text{cyc}}$ move rows only between SUPPORTED and WEAK; they leave every column of this table unchanged by construction, and their effect on the SUPPORTED share and the false-SUPPORTED rate is reported in Appendix K.

| Setting | $\text{Avg}_{14}$ | UCR | aUCR | Cert.$F_1$ | Coverage | Gap | Latency |
|---|---|---|---|---|---|---|---|
| $K = 2$ | 85.10 | 7.6 | 5.6 | 75.4 | 88.2 | 0.052 | 1.14 |
| $K = 4$ | 86.26 | 5.9 | 4.0 | 80.1 | 91.7 | 0.036 | 1.16 |
| $K = 8$ (default) | **86.94** | 4.9 | 2.5 | 84.2 | 94.1 | 0.027 | 1.18 |
| $K = 16$ | 86.98 | 4.7 | 2.4 | 84.6 | 94.4 | 0.025 | 1.25 |
| $T = 5$ | 85.65 | 6.5 | 4.7 | 78.2 | 90.0 | 0.047 | 1.15 |
| $T = 10$ | 86.42 | 5.5 | 3.3 | 82.1 | 92.7 | 0.033 | 1.17 |
| $T = 25$ | 87.01 | 4.7 | 2.4 | 84.5 | 94.3 | 0.026 | 1.21 |
| $\tau_{\varnothing}$ $-25\%$ / $+25\%$ | 86.94 | 4.9 | 2.2 / 3.1 | 82.7 / 81.6 | 92.4 / 95.3 | 0.027 | 1.18 |
| $\alpha$ ($\Delta_{\text{led}}$) $\times 0.5$ / $\times 2$ | 86.73 | 5.3 / 5.0 | 2.8 / 2.6 | 82.6 / 84.0 | 93.6 / 94.0 | 0.031 / 0.024 | 1.18 |
| $\beta$ ($\mathcal{L}_{\text{cf}}$) $\times 0.5$ / $\times 2$ | 86.76 | 5.4 / 4.9 | 3.0 / 2.5 | 81.9 / 84.1 | 93.4 / 94.1 | 0.031 / 0.027 | 1.18 |
| $\gamma$ ($\mathcal{L}_{\text{ver}}$) $\times 0.5$ / $\times 2$ | 86.89 | 5.0 / 4.9 | 2.6 / 2.5 | 83.6 / 84.1 | 94.0 / 94.1 | 0.027 / 0.027 | 1.18 |
| $\eta$ ($\mathcal{L}_{\text{cal}}$) $\times 0.5$ / $\times 2$ | 86.91 | 4.9 / 5.0 | 2.5 / 2.6 | 84.0 / 83.9 | 94.1 / 94.0 | 0.027 / 0.027 | 1.18 |
| $\lambda_{\varnothing}$ $\times 0.5$ / $\times 2$ (stop-gradient term; no training effect) | 86.94 | 4.9 / 4.9 | 2.5 / 2.5 | 84.2 / 84.2 | 94.1 / 94.1 | 0.027 / 0.027 | 1.18 |
| $\lambda_g$ ($\Gamma$) $\times 0.5$ / $\times 2$ | 86.82 | 5.2 / 4.9 | 2.8 / 2.5 | 83.1 / 84.1 | 93.8 / 94.1 | 0.030 / 0.027 | 1.18 |
| $\lambda_s$ ($\Omega_s$) $\times 0.5$ / $\times 2$ | 86.64 | 5.6 / 5.1 | 3.2 / 2.7 | 81.5 / 83.6 | 93.4 / 94.0 | 0.034 / 0.028 | 1.18 |

Table G.18: **Null share does not track evidence count.** Mean null allocation $n^{\text{alloc}}$ and diffuse share $\bar{\delta}$ per claim, the ledger's NULL rate, and strict hUCR on the 520-prompt audit, sliced by evidence count $n$ relative to claim count $\ell$, by $\ell$, and by input type (model-derived quantities, the shares and the NULL rate, are five-seed means with seed SD $\leq 0.003$ and $\leq 0.3$ pp; hUCR is measured on the audited seed-17 outputs). Because concrete atoms are capacity bounds and only the null row is free, unused capacity on an evidence-rich input is not null mass: the null share is 0.036 on documents with $n > 100$ atoms and 0.044 on photographs with $n \leq 12$ (Spearman correlation between $n$ and $n^{\text{alloc}}$ across examples $-0.04$), and the diffuse share stays near 0.03 because irrelevant candidates receive $e^{-\Delta D/\varepsilon} \approx 10^{-3}$ of a column. Slices weight back to the pooled values (0.041 / 0.031 / 5.9% / 4.6%). The uniform-capacity variant $u \equiv 1$ of Table G.11 gives 0.043 / 0.033.

| Slice | share of examples (%) | mean $n$ | mean $\ell$ | $n^{\text{alloc}}$ | $\bar{\delta}$ | NULL rate (%) | hUCR (%) |
|---|---|---|---|---|---|---|---|
| All audited examples | 100 | 23 | 3.0 | 0.041 | 0.031 | 5.9 | 4.6 |
| $n \geq 3\ell$ (evidence-rich) | 71 | 29 | 3.0 | 0.038 | 0.034 | 5.4 | 4.3 |
| $n < 3\ell$ (evidence-poor) | 29 | 8 | 3.0 | 0.046 | 0.024 | 7.1 | 5.3 |
| $\ell = 1$ | 18 | 21 | 1.0 | 0.035 | 0.029 | 4.9 | 3.8 |
| $\ell = 2$–3 | 54 | 23 | 2.6 | 0.040 | 0.031 | 5.7 | 4.5 |
| $\ell \geq 4$ | 28 | 25 | 5.1 | 0.047 | 0.033 | 6.9 | 5.4 |
| Documents ($n > 100$) | 6 | 164 | 3.4 | 0.036 | 0.041 | 5.2 | 4.0 |
| Photographs ($n \leq 12$) | 31 | 9 | 2.7 | 0.044 | 0.022 | 6.6 | 5.0 |

by annotator recall in Appendix N; the null-dominance rule alone accounts for 16 of the 35 NULL hits (16/62, precision 26%, recall 23%) and the contradiction flag for the other 19 (19/27), so the localisation the certificate offers rests on both mechanisms. In deployment the distinction that carries information is therefore two-level, asserted against NULL, with WEAK retained as a low-confidence marker for annotators. A row becomes WEAK by one of two routes (Algorithm 3): a score below the threshold, $\hat{s}_j < \hat{\tau}_{\text{sup}}$, or a residual above it, $r_j > \tau_{\text{cyc}}$. On the audit the two routes are very unequal: 12 of the 392 WEAK rows are WEAK by score (5 adjudicated unsupported) and 380 are WEAK by residual with $\hat{s}_j \geq \hat{\tau}_{\text{sup}}$ (6 unsupported, 1.6%), so the residual gate removes rows that are almost as reliable as the SUPPORTED rows; this is the operational meaning of the statement that $r_j$ measures route disagreement and not unfunded content, and it is why the two WEAK rows shown in the figures (Figures A.10 and J.12), both WEAK by score with $n_j^{\text{alloc}} < \tau_{\varnothing}$ and no contradiction flag, are the rarer kind. The residual distribution is right-skewed: $r_j$ exceeds $\tau_{\text{cyc}} = 0.05$ on 27% of asserted rows, which come from the 31% of audited answers whose $\Delta_{\text{cyc}}$ exceeds 0.05 (median $\Delta_{\text{cyc}}$ over answers 0.008, mean 0.019), while every row shown in the galleries with a SUPPORTED verdict has $r_j \leq 0.031$ by construction. Lowering $\tau_{\text{cyc}}$ therefore trades SUPPORTED coverage for no measurable gain in precision on this data, and raising it to 0.10 would move 214 of the 380 residual-gated rows

Table G.19: **Standard generative-hallucination benchmarks and relabelled POPE.** Long-form descriptions are produced under a captioning prompt ("Describe the image in detail.", greedy, at most 512 tokens), not the concise prompt of the main evaluation, so the fine-tuning baselines are evaluated in the setting for which they were developed. Object HalBench: $\text{CHAIR}_s$ / $\text{CHAIR}_i$ (%, lower is better) on the 300-image split used by RLHF-V; AMBER generative task: CHAIR, Cover, Hal, Cog (Cover higher is better; the others lower); MMHal-Bench: GPT-4-judged score (0–6, higher) and hallucination rate (lower). RePOPE re-scores the adversarial POPE split against corrected labels (Neuhaus and Hein, 2025), which reverse a substantial share of the original positive labels and remove ambiguous items; $\Delta F_1$ is the change from the original labels, so a system that answers "yes" on the relabelled positives loses more. Every system loses $F_1$ under relabelling; the ledger loses least and keeps the largest margin. Trained systems are three-seed means (five for the ledger); the backbone is a single run.

| | Object HalBench | | AMBER (generative) | | | | MMHal-Bench | | RePOPE | |
|---|---|---|---|---|---|---|---|---|---|---|
| System | $\text{CHAIR}_s\downarrow$ | $\text{CHAIR}_i\downarrow$ | CHAIR↓ | Cover↑ | Hal↓ | Cog↓ | Score↑ | Hal. rate↓ | $F_1\uparrow$ | $\Delta F_1$ vs. POPE |
| Qwen3-VL-235B (backbone) | 6.8 | 3.6 | 3.9 | 52.1 | 20.4 | 2.1 | 3.62 | 0.28 | 90.3 | −3.6 |
| + yes/no logit-bias correction | – | – | – | – | – | – | – | – | 90.9 | −4.1 |
| + tool outputs in context | 5.2 | 2.7 | 3.1 | 52.4 | 16.5 | 1.6 | 3.80 | 0.24 | 91.2 | −4.0 |
| + LURE | 5.6 | 3.0 | 3.4 | 51.6 | 18.1 | 1.9 | 3.71 | 0.26 | 91.0 | −4.3 |
| + HA-DPO | 5.1 | 2.7 | 3.1 | 51.7 | 16.8 | 1.7 | 3.78 | 0.24 | 91.4 | −4.0 |
| + POVID | 4.9 | 2.6 | 3.0 | 51.8 | 16.2 | 1.6 | 3.82 | 0.23 | 91.6 | −4.0 |
| + RLAIF-V | 4.4 | 2.3 | 2.7 | 51.8 | 14.9 | 1.4 | 3.90 | 0.21 | 92.2 | −3.9 |
| + on-policy grounder-DPO | 4.2 | 2.1 | 2.6 | 52.0 | 14.1 | 1.3 | 3.94 | 0.20 | 92.6 | −3.7 |
| + on-policy GRPO | 4.0 | 2.0 | 2.5 | 52.2 | 13.6 | 1.3 | 3.97 | 0.19 | 92.7 | −3.7 |
| + tool outputs in context + grounder-DPO | 4.0 | 2.0 | 2.5 | 52.3 | 13.4 | 1.2 | 3.98 | 0.19 | 92.9 | −3.6 |
| Control A | 4.5 | 2.3 | 2.8 | 51.9 | 15.0 | 1.4 | 3.88 | 0.21 | 92.3 | −3.8 |
| Control B1 | 3.4 | 1.7 | 2.1 | 52.5 | 11.4 | 1.1 | 4.01 | 0.18 | 94.1 | −2.9 |
| LEDGER-VLM | **3.1** | **1.6** | **2.0** | **52.6** | **10.8** | **1.0** | **4.05** | **0.17** | **94.7** | **−2.7** |

Table G.20: **Where capacities bind.** Strict verifier UCR of the ledger, of the ledger without its row cap (the penalty-matched control, since B1 and B2 lack the row-degree component of $\Omega_s$), and of Controls B1 and B2 on (i) the evaluation prompts that contain at least one binding evidence row, (ii) the complementary slack prompts, and (iii) a constructed capacity-binding slice of 600 prompts (200 counting questions whose answer names a count larger than the number of detected instances, 200 images with duplicated objects of one class, 200 document lines that three or more claims cite). The binding/slack split is decided once, at the $T = 200$ fixed point of the seed-17 full ledger's plans, over all 39,086 evaluation prompts of Table C.2; 2,384 prompts (6.1%) contain a binding row. Paired differences are B1 minus the ledger and B1 minus the no-row-cap ledger (three seeds for the controls; prompt-clustered 95% intervals for B1 − ledger are $[0.6, 3.0]$, $[-0.3, 1.1]$, $[1.1, 4.6]$, $[0.6, 3.9]$, $[0.7, 3.8]$, and $[1.4, 3.5]$ pp for the six rows; for B1 − no-row-cap they are $[0.2, 2.4]$, $[-0.9, 0.3]$, $[0.4, 2.9]$, $[0.2, 2.7]$, $[0.2, 2.6]$, and $[0.6, 2.4]$). Against the penalty-matched control, coupling gains 1.3 pp on binding prompts and 1.5 pp on the slice and loses 0.3 pp on slack prompts; the slice is not part of any headline number. Under demand-scaled capacities (Table G.11), which make rows bind on 38% of prompts, the ledger's margin over B1 on the full evaluation set is 1.1 pp while its own UCR is 5.1%.

| Subset | prompts | LEDGER-VLM | no row cap | Control B1 | Control B2 | B1 − ledger | B1 − no row cap |
|---|---|---|---|---|---|---|---|
| Binding rows present (evaluation set) | 2,384 | **6.1** | 6.6 | 7.9 | 8.8 | +1.8 | +1.3 |
| Slack rows only (evaluation set) | 36,702 | **4.8** | 5.5 | 5.2 | 6.1 | +0.4 | −0.3 |
| Constructed slice: counting | 200 | **8.2** | 9.4 | 11.0 | 12.1 | +2.8 | +1.6 |
| Constructed slice: duplicated objects | 200 | **6.9** | 7.7 | 9.1 | 9.8 | +2.2 | +1.4 |
| Constructed slice: shared OCR line | 200 | **7.1** | 7.9 | 9.3 | 10.0 | +2.2 | +1.4 |
| Constructed slice, pooled | 600 | **7.4** | 8.3 | 9.8 | 10.6 | +2.4 | +1.5 |

Table G.21: Qwen3-VL $6 \times 6$ scaling sweep, one run per cell (seed 17): $\text{Avg}_{\text{scal}}$ / certificate $F_1$ / strict UCR, where $\text{Avg}_{\text{scal}}$ is the mean of the five tasks MMMU, ChartQA, AI2D, MMBench, RefCOCOg (the ledger's five-seed mean over these tasks in Table G.10 is 90.70; the backbone's is 89.36). Backbones are the released Qwen3-VL checkpoints; the token axis is the number of ledger-training tokens processed (visual tokens included). The 235B-A22B / 4.9B cell is the seed-17 run of the main configuration (90.62 / 83.96 / 5.23) and differs from the five-seed means by seed noise only. Four cells (2B/0.15B, 8B/1.2B, 32B/4.9B, 235B/0.6B) were replicated with three seeds: $F_1$ SDs 0.12–0.21, UCR SDs 0.11–0.19.

| Backbone ↓ / ledger-training tokens → | 0.15B | 0.3B | 0.6B | 1.2B | 2.4B | 4.9B |
|---|---|---|---|---|---|---|
| Qwen3-VL-2B | 72.41/61.71/12.28 | 73.15/66.87/11.21 | 73.56/70.40/10.07 | 73.85/73.41/9.54 | 74.17/76.59/8.92 | 74.58/78.41/8.45 |
| Qwen3-VL-4B | 76.71/63.49/11.71 | 77.53/68.22/10.41 | 78.06/71.65/9.62 | 78.34/74.83/8.83 | 78.73/77.91/8.27 | 78.77/80.33/7.74 |
| Qwen3-VL-8B | 80.64/65.26/11.02 | 81.56/69.54/9.87 | 81.80/73.56/8.90 | 82.12/76.37/8.14 | 82.34/78.87/7.28 | 82.58/81.25/6.89 |
| Qwen3-VL-30B-A3B | 82.13/65.60/10.77 | 82.60/70.86/9.68 | 83.11/73.86/8.77 | 83.57/76.69/7.79 | 83.66/79.33/7.22 | 83.85/81.79/6.71 |
| Qwen3-VL-32B | 85.92/67.29/9.96 | 86.28/71.78/8.88 | 86.88/75.32/7.92 | 87.07/78.29/7.33 | 87.40/80.78/6.68 | 87.63/82.98/6.07 |
| Qwen3-VL-235B-A22B | 89.42/70.03/9.56 | 89.79/73.90/8.37 | 90.22/77.38/7.36 | 90.14/80.13/6.46 | 90.35/81.91/5.73 | **90.62/83.96/5.23** |

Table H.22: Full five-regime UCR stress sweep (as-emitted strict UCR over non-refused answers, verifier grader, seven severities; five seeds for the ledger, three for on-policy grounder-DPO, the reference baseline of the stress tests; the backbone is a single greedy run per severity). Severity 0 is the unperturbed evaluation and reproduces the main-result values (11.4 / 7.6 / 4.9). Refused answers (generator refusals only; no policy is applied) contribute no claims; Table H.23 reports the refusal rate and the task success, which does penalise refusal. Seed SDs grow with severity: 0.10–0.24 pp for the ledger, 0.18–0.41 pp for grounder-DPO.

| Regime | System | 0 | 0.15 | 0.30 | 0.45 | 0.60 | 0.75 | 0.90 |
|---|---|---|---|---|---|---|---|---|
| Conflict | Qwen3-VL | 11.40 | 15.54 | 20.32 | 25.76 | 31.26 | 37.08 | 43.10 |
| Conflict | + grounder-DPO | 7.60 | 10.54 | 14.08 | 18.18 | 22.22 | 26.40 | 30.90 |
| Conflict | LEDGER-VLM | **4.90** | **5.98** | **7.50** | **8.92** | **10.60** | **12.41** | **14.10** |
| Distractor | Qwen3-VL | 11.40 | 15.28 | 19.87 | 25.11 | 30.23 | 36.20 | 41.60 |
| Distractor | + grounder-DPO | 7.60 | 10.56 | 14.05 | 17.74 | 21.67 | 26.00 | 30.20 |
| Distractor | LEDGER-VLM | **4.90** | **6.07** | **7.29** | **8.80** | **10.57** | **11.92** | **13.60** |
| OCR noise | Qwen3-VL | 11.40 | 15.12 | 19.39 | 24.21 | 29.17 | 34.38 | 40.20 |
| OCR noise | + grounder-DPO | 7.60 | 10.43 | 13.83 | 17.55 | 21.11 | 25.48 | 29.50 |
| OCR noise | LEDGER-VLM | **4.90** | **5.96** | **7.08** | **8.56** | **10.26** | **11.62** | **13.10** |
| Modality drop | Qwen3-VL | 11.40 | 16.43 | 22.32 | 28.89 | 35.79 | 42.94 | 50.30 |
| Modality drop | + grounder-DPO | 7.60 | 11.36 | 15.94 | 20.95 | 26.23 | 31.88 | 37.40 |
| Modality drop | LEDGER-VLM | **4.90** | **6.59** | **8.55** | **10.90** | **13.09** | **15.41** | **17.70** |
| Temporal | Qwen3-VL | 11.40 | 15.91 | 21.40 | 27.19 | 33.20 | 40.53 | 46.70 |
| Temporal | + grounder-DPO | 7.60 | 11.02 | 14.87 | 19.28 | 23.82 | 28.61 | 33.40 |
| Temporal | LEDGER-VLM | **4.90** | **6.20** | **7.73** | **9.32** | **11.56** | **12.94** | **15.00** |

Table H.23: Populations behind the stress numbers at severity 0.75 (seed SDs 0.28–0.56 for task success, 0.14–0.50 pp for UCR, 0.19–0.47 pp for refusal). *Task success* is measured over all prompts and counts a refusal as a failure; *accuracy on non-refused* is task success divided by the non-refused share; *UCR* is computed over the parsed claims of non-refused answers; the *refusal rate* counts answers whose decoded text declines to answer. Every refusal is a generator refusal; no deployment policy is applied to any reported number, and NULL rows are flags on claims that remain in the scored answer. The ledger's refusal rate is within 1.2 pp of the backbone's under distractor, OCR-noise, and temporal perturbation, 3.8 pp higher under contradictory retrieval, and 8.4 pp higher under modality drop, where the evidence is absent; under modality drop this refusal costs task success and the ledger's advantage shrinks to 1.1. Accuracy on non-refused answers stays below every system's clean $\mathrm{Avg}_{14}$.

| Regime | System | Task success↑ (all prompts) | Accuracy on non-refused | UCR↓ (non-refused) | Generator refusal rate |
|---|---|---|---|---|---|
| Conflict | Qwen3-VL / grounder-DPO / LEDGER-VLM | 69.9 / 73.6 / **74.4** | 76.1 / 80.3 / 84.4 | 37.08 / 26.40 / **12.41** | 8.1 / 8.4 / 11.9 |
| Distractor | Qwen3-VL / grounder-DPO / LEDGER-VLM | 72.0 / 75.1 / **78.6** | 77.8 / 81.5 / 85.5 | 36.20 / 26.00 / **11.92** | 7.5 / 7.9 / 8.1 |
| OCR noise | Qwen3-VL / grounder-DPO / LEDGER-VLM | 69.4 / 72.9 / **76.2** | 76.1 / 80.1 / 84.1 | 34.38 / 25.48 / **11.62** | 8.8 / 9.0 / 9.4 |
| Modality drop | Qwen3-VL / grounder-DPO / LEDGER-VLM | 61.7 / 64.3 / **65.4** | 68.7 / 72.2 / 80.3 | 42.94 / 31.88 / **15.41** | 10.2 / 11.0 / 18.6 |
| Temporal | Qwen3-VL / grounder-DPO / LEDGER-VLM | 70.3 / 73.4 / **76.5** | 77.9 / 81.5 / 85.6 | 40.53 / 28.61 / **12.94** | 9.7 / 9.9 / 10.6 |

Table K.24: Confusion of the ledger's three verdicts against blinded human adjudication on the 1,512 adjudicable claims of the 520-prompt audit (NULL rows are flags; every claim stays in the scored answer). Precision and recall treat "non-SUPPORTED" as the detector of unsupported claims. Most NULL over-flags are claims whose evidence was never atomised (61% of the 54 supported NULL rows) and are the price of the 97.7% precision of SUPPORTED. Of the 89 NULL rows, 27 were produced by the contradiction flag (19 adjudicated unsupported, precision 70%) and 62 by null dominance; without the flag, 14 of those 19 and 6 of the 8 supported flag-driven rows would have been stamped SUPPORTED (the rest fail the score or residual test), and the test false-SUPPORTED rate would have been $38/1{,}051 = 3.6\%$, above $\vartheta$. Strict criterion throughout.

| Verdict | rows | adjudicated supported | adjudicated unsupported | share of rows |
|---|---|---|---|---|
| SUPPORTED | 1,031 | 1,007 | 24 | 68.2% |
| WEAK | 392 | 381 | 11 | 25.9% |
| NULL | 89 | 54 | 35 | 5.9% |

SUPPORTED precision $1{,}007/1{,}031 = 97.7\%$ (false-SUPPORTED rate 2.3%); non-SUPPORTED recall of unsupported claims $46/70 = 65.7\%$, precision $46/481 = 9.6\%$; NULL recall $35/70 = 50.0\%$, precision $35/89 = 39.3\%$; claim coverage (SUPPORTED or WEAK) $1{,}423/1{,}512 = 94.1\%$; asserted-claim hUCR $35/1{,}423 = 2.5\%$.

to SUPPORTED at an audited false-SUPPORTED rate of 2.4%; we keep $\tau_{\mathrm{cyc}} = 0.05$ as selected before the audit.

**Scorer ablation on shared answers.** The "without plan features" scorer of Table 3 replaces $\varphi_j$ by the verifier logit $u_j$ alone, with its own logistic and isotonic maps fit on the verifier-fit split and its own learn-then-test threshold; it isolates what the explicit plan statistics ($\bar{c}_j$, $r_j$, $n_j^{\mathrm{alloc}}$, the maximal share, and $\tilde{H}_j$) add; it is not a transport-free scorer, because $v_\omega$ itself reads the column $\Pi^{\mathrm{EC}}_{\cdot,j}$, so the comparison bounds the value of the explicit statistics and not of the plan as a whole. On control

Table K.25: Where the HallusionBench gain comes from: rate of answering "yes" on questions whose gold answer is "no" (yes-bias), accuracy on questions whose gold answer is "yes", and consistency across the paired question sets, for the additive and attribution controls of Table G.11, the on-policy baselines and RLAIF-V, and a yes/no logit-bias correction of the backbone tuned on held-out data, which bounds how much of the gain a pure prior shift can deliver (its POPE $F_1$ is 95.0 against the ledger's 97.4). Trained rows are seed means (five seeds for the ledger; three otherwise); the backbone is a single greedy run. Hall.Acc is qAcc.

| System | yes-bias on "no" items (%) ↓ | accuracy on "yes" items (%) | pair consistency (%) ↑ | Hall.Acc |
|---|---|---|---|---|
| Qwen3-VL-235B (backbone) | 42.4 | 87.9 | 51.0 | 64.2 |
| Answer-only LoRA | 41.8 | 88.1 | 51.4 | 65.0 |
| $+\mathcal{L}_{\mathrm{ul}}$ only | 26.3 | 87.5 | 57.9 | 69.4 |
| $+\Delta_{\mathrm{cyc}}$ only via $\Pi^{\mathrm{EA}}$ | 30.8 | 87.6 | 56.3 | 68.2 |
| + InfoNCE alignment only | 30.6 | 87.6 | 56.4 | 68.5 |
| $+\mathcal{L}_{\mathrm{ul}} + \Delta_{\mathrm{cyc}}$ | 23.1 | 87.2 | 60.4 | 71.0 |
| $+\mathcal{L}_{\mathrm{cf}}$ only | 28.1 | 87.3 | 57.2 | 69.1 |
| $+\Delta_{\mathrm{led}}$ only | 30.4 | 87.4 | 56.1 | 68.7 |
| Control A (grounder-weighted $\mathcal{L}_{\mathrm{ul}}$) | 25.6 | 87.4 | 58.2 | 69.8 |
| Control B1 (softmax + dustbin, $\Pi^{\mathrm{EA}}$ kept) | 20.4 | 86.9 | 62.3 | 73.1 |
| Control B2 (softmax + dustbin, no $\Pi^{\mathrm{EA}}$) | 22.5 | 87.1 | 60.7 | 71.6 |
| Control C (InfoNCE in place of $\Delta_{\mathrm{cyc}}$) | 21.0 | 86.9 | 61.9 | 72.8 |
| RLAIF-V | 25.0 | 86.9 | 58.9 | 69.5 |
| Backbone + yes/no logit-bias correction (held-out tuned) | 34.6 | 87.4 | 53.9 | 67.0 |
| On-policy grounder-DPO | 23.8 | 87.0 | 59.6 | 70.4 |
| On-policy GRPO | 23.2 | 87.1 | 60.1 | 70.9 |
| Grounder-DPO $+\Delta_{\mathrm{cyc}}$ / $+\Delta_{\mathrm{cyc}} + \mathcal{L}_{\mathrm{cf}}$ | 22.3 / 21.3 | 87.0 / 86.9 | 60.9 / 61.7 | 71.9 / 72.6 |
| Grounder-DPO $+\mathcal{L}_{\mathrm{ul}}$ / all three | 22.0 / 20.0 | 87.0 / 86.9 | 61.2 / 62.5 | 72.1 / 73.4 |
| LEDGER-VLM | **19.0** | 86.8 | **63.7** | **74.6** |

Table K.26: HallusionBench under its three official metrics, by subset and question type, and POPE $F_1$ with the yes-ratio (five-seed means with SD for the ledger; the backbone is a single run). Hall.Acc in every other table is qAcc. The "no"-gold and "yes"-gold rows are per-question accuracies and combine (45/55 weighting) to aAcc. The ledger's yes-ratio moves from the backbone's 53.1% to 48.9% around the 50.0% of the original labels; because those labels contain annotation errors, the relabelled RePOPE margin of Table G.19 is the object-hallucination evidence we rely on, and POPE $F_1$ is reported for comparability with prior work. **Contamination check**: images and questions of the evaluated HallusionBench and POPE slices are disjoint from the 1.8M training mixture, the 46k records, and the risk-control sets (exact-match and perceptual-hash check at Hamming distance $\leq 4$ over 1.8M images; 0 hits; the risk-control carve-outs of Table C.2 are removed from the evaluated slice for every system). The human audit samples its prompts from the evaluated slices.

| Subset | backbone | LEDGER-VLM | Δ |
|---|---|---|---|
| qAcc (question-pair; Hall.Acc) | 64.2 | 74.6 (0.44) | +10.4 |
| aAcc (per question) | 74.2 | 84.2 (0.35) | +10.0 |
| fAcc (per figure) | 45.3 | 56.1 (0.58) | +10.8 |
| Visual dependent (VD), qAcc | 62.1 | 73.4 (0.44) | +11.3 |
| Visual supplement (VS), qAcc | 66.8 | 76.0 (0.39) | +9.2 |
| "No"-gold questions, accuracy | 57.6 | 81.0 (0.51) | +23.4 |
| "Yes"-gold questions, accuracy | 87.9 | 86.8 (0.36) | −1.1 |
| Figure / chart questions, qAcc | 60.4 | 71.9 (0.48) | +11.5 |
| Map / table questions, qAcc | 67.2 | 76.7 (0.41) | +9.5 |
| POPE $F_1$ | 93.9 | 97.4 (0.12) | +3.5 |
| POPE yes-ratio (gold 50.0) | 53.1 | 48.9 (0.18) | −4.2 |

and backbone answers the plan features add 1.0 pp of coverage at controlled risk; the verifier head accounts for the remaining 0.8–1.1 pp of the margin over the classifier, which it owes to attending over the candidate atoms jointly rather than scoring the claim against each atom independently.

**Representation alignment.** The mean cosine between an answer-unit state $z_k^A$ and the embeddings of the atoms that fund it is 0.62 for the full ledger, 0.41 without $\Pi^{\mathrm{EA}}$, and 0.38 for answer-only LoRA, consistent with $\Delta_{\mathrm{cyc}}$ acting on the adapters through $z^A$; this is offered as a plausible second channel for the task-score gains, not as an established mechanism.

## L FALSIFICATION MATRIX

Table L.27 lists, for each core claim, the pre-specified measurement that could have weakened it and, where the measurement produced a qualified result, what was observed; the qualified rows are the ones a reader should weigh against the claims of Section 5.4.

Table L.27: Pre-specified tests of each core claim: the measurement that would have weakened the claim, and the outcome where the test produced a qualified result.

| Core claim | Decisive measurement | Expected falsifying result, or observed qualification (*Outcome*) |
|---|---|---|
| The objective reduces unsupported content | hUCR, UCR, HallusionBench, POPE, long-form benchmarks, RePOPE | Human-graded UCR fails to fall, gains appear only at low coverage, or long-form and relabelled benchmarks disagree |
| Null source exposes missing evidence | Modality-drop UCR, ME-AUC | Removing the null source leaves UCR unchanged |
| The two-route consistency term adds to its nearest controls | B1 vs. B2 (paired), Control C, portability rows | *Outcome*: removing the direct route from the softmax construction costs 0.9 pp, with a seed-paired interval of $[0.3, 1.5]$, and replacing $\Delta_{\mathrm{cyc}}$ by InfoNCE costs 0.7 pp, interval $[0.2, 1.2]$; the increments are consistent in sign across the paired tests but modest next to the unlikelihood and swap terms |
| Coupling across claims matters | Proposition 1; binding fraction; Table G.20 | *Outcome*: capacities bind on 6% of examples; B1 is within 0.5 pp overall and behind by 1.8 pp only on the binding subset |
| Certificates are counterfactually consistent | Counterfactual swaps and masking | Funding sets unchanged after decisive evidence is swapped or masked |
| Selective risk control holds | Test false-SUPPORTED rate at $\hat{\tau}_{\mathrm{sup}}$; coverage at $\vartheta$ | Test rate above $\vartheta = 3\%$ in-distribution, or coverage at controlled risk no better than the controls trained on the same annotations |
| Training, not emission, separates ledger from grounder | Same-annotation controls; additive controls | Post-hoc grounder or claim classifier trained on the 6.2k annotations matches the ledger's UCR |
| The coupled solve contributes beyond its ingredients | Controls A, B1, B2 (Table 3) | *Outcome*: Control B1 (softmax + dustbin with the direct route) is within 0.5 pp of the ledger, an audited difference whose interval includes zero (Section 5.4); Controls A ($+2.3$) and B2 ($+1.4$) are distinguishable from the ledger, Control C ($+0.7$) only under the seed-level test. On this mixture the coupling across claims contributes on the binding subset and little elsewhere; the largest single-term effects are the unlikelihood and swap terms |
| Null term should be trained through | Degenerate variant ($\Delta_{\mathrm{null}}$ back-propagated) | *Outcome*: null share falls to 0.012 while UCR rises to 7.6%; the gradient is removed and $\Delta_{\mathrm{null}}$ is a diagnostic only |
| NULL localises unsupported claims | NULL recall / precision (Table K.24) | *Outcome*: recall 50%, precision 39%; NULL is a pointer, not a detector |
| WEAK is a distinct risk level | Error rate of WEAK vs. SUPPORTED rows | *Outcome*: 2.8% vs. 2.3%, with 380 of 392 WEAK rows gated by the residual rather than the score (Appendix K); the two-level view (asserted / NULL) carries almost the same information |
| Diagnostics track risk | 35-cell stress correlation | Near-zero diagnostics with persistently high J-ind UCR |
| Mechanism is not backbone-specific | InternVL3.5-38B replication | Relative UCR reduction below half of the Qwen3 value |

# M DEPLOYMENT POLICY

The certificate never edits an answer; the table below is the recommended mapping from certificate rows to product actions, including the 10% spot-check of SUPPORTED rows motivated by Appendix N.

**Optional ledger-constrained decoding (excluded from the main results).** A deployment may let the accounting act at inference: on the 520 audit prompts, re-decoding the 14.2% of answers that contain a NULL row with those spans masked from the draft lowered strict hUCR from 4.6% to 3.9% at $1.9\times$ latency on the affected prompts, with $\mathrm{Avg}_{14}$ unchanged ($\pm 0.1$). We report it to show that the ledger can constrain generation when a product wants it to; the headline numbers keep generation and certification separate.

Table M.28: Certificate-to-action policy.

| Case | Certificate condition | Action |
|---|---|---|
| Accept | All high-risk rows SUPPORTED; $n^{\mathrm{alloc}} \le \tau_{\varnothing}$ and $r \le \tau_{\mathrm{cyc}}$ | Return answer with compact certificate |
| Annotate | Low-risk row is WEAK or NULL with no contradiction flag | Return the full answer; show the row's verdict and funding atoms beside it (the text is not edited) |
| Reject | High-risk row is NULL, or coverage below budget | Refuse that claim; request missing evidence |
| Escalate | Contradiction flag on a high-mass row | Route to human audit with the claim and the contradicting funding atom displayed (a deployment action, not a verdict) |
| Monitor | Drift in UCR, ME-AUC, or Cert.$F_1$ beyond tolerance | Refresh the labeled audit sample and recalibrate |

## N HUMAN STUDY PROTOCOL

Fourteen trained internal annotators (not authors) participated under an institutional minimal-risk exemption; written consent, \$25/hour compensation, skip and withdrawal rights, no personal data. Auditors saw paired answers for the same prompt in randomised order with identities hidden; certificate-bearing answers exposed only row verdicts, evidence summaries, and residual indicators. Each of eight comparisons used 200 prompts across document, chart, hallucination-diagnostic, video, and grounding families. Every prompt in every arm was judged by three annotators, with the assignment balanced so that each annotator saw about 340 items in total. The win rate on a criterion is the share of the 600 judgements in the arm, three per prompt, that favour the ledger side, so it moves in steps of one sixth of a percentage point; its interval is a prompt-clustered bootstrap over the 200 prompts with 10,000 resamples, which accounts for the correlation among the three judgements on one prompt. The per-prompt verification time is the median of the three annotators' times; time changes are relative differences of mean per-prompt times, with prompt-level bootstrap intervals stratified by annotator so that repeated measurements by the same annotator are not treated as independent. Per-annotator logs were deleted after aggregation. **Same-answer arm.** In the decisive comparison both sides show the *same* ledger answer; one side carries the ledger certificate and the other a certificate produced by the grounder trained on the same annotations, run on that answer. Both are rendered in one schema (claim text, funding atom identifiers with thumbnails, one support bar carrying $\hat{s}_j$ or the grounder's support probability, verdict); the residual, share, and contradiction fields that only the ledger produces are hidden on both sides, so the two displays differ only in which atoms are cited, how long the bar is, and which verdict is stamped. Asked after each item which system produced the certificate, annotators were right in 312 of the 600 judgements, 52%, against a chance level of 50%. **No-certificate arm.** The ledger answer is shown without any certificate against the backbone answer, which bounds how much of the win rate is due to the answer text alone.

**What the comparisons can show.** Against systems that show no certificate, higher trust and auditability are partly built into the design of the study, so those four comparisons mainly bound the cost of the interface. The comparisons against the post-hoc grounder and the claim-classifier control show a certificate on both sides but different answers, so they confound answer and certificate. The same-answer arm removes the confound: with the answer fixed, the ledger certificate wins $64.8/70.2/68.5\%$ of judgements and saves $19.7\%$ of verification time, and auditors find $81\%$ of adjudicated-unsupported claims with it against $69\%$ with the grounder certificate on the same answer. The no-certificate arm ($58.3/57.2/54.8\%$, time $+4.1\%$) shows that the answer text alone accounts for a small part of the win rate. Certificate value therefore has two measured components: the answer ($\approx 8$ pp over parity) and the certificate rows ($\approx$ 15–20 pp).

**External replication.** Because the main panel consisted of internal (non-author) annotators, six annotators recruited through a crowd-work vendor and screened with a 20-item qualification test repeated the study on a 60-prompt subsample per comparison. Win rates against Qwen3-VL were $79.5/84.2/83.0\%$ (helpfulness / trust / auditability) and the verification-time change was $-35.1\%$; against post-hoc grounding, $72.8/76.1/74.6\%$ and $-26.4\%$; in the same-answer arm, $62.1/67.5/65.0\%$ and $-17.9\%$; every value lies within 4 pp of the internal panel.

**Verification accuracy.** To assess verification accuracy alongside speed, 120 prompts per condition were seeded with adjudicated-unsupported claims from the audit and auditors were asked to mark every claim they considered unsupported. With certificates auditors marked $78.3\%$ of the seeded claims (recall) at $91.2\%$ precision, against $61.4\%$ recall and $88.7\%$ precision without certificates. Recall was $84\%$ on NULL rows, $71\%$ on WEAK rows, and $22\%$ on the false-SUPPORTED rows, because auditors reported skipping rows stamped SUPPORTED. The deployment policy of Appendix M therefore adds a random $10\%$ spot-check of SUPPORTED rows, which on this sample lifts recall on false-SUPPORTED rows to $34\%$ at a $6\%$ increase in verification time.

## O BROADER IMPACT AND CORNER CASES

Ledgers are intended to reduce unsupported multimodal claims in scientific, document, and safety-inspection settings. Risks: over-trust when the evidence atomiser misses decisive content; user frustration when row-level refusal fires on evidence obvious to a human; uneven coverage across

Table N.29: MMMU accuracy by category on the 900-question validation split. The backbone is a single frozen run and its category accuracies are exact counts; the ledger's per-seed correct counts are listed for seeds $17/29/41/53/67$, and its mean and sample SD are computed from those integer counts, so every value can be re-derived from the table; $\Delta$ is the difference of the displayed means. The size-weighted means reproduce the $79.6$ and $81.2$ of Table G.10, and the SD of the per-seed totals is the MMMU cell of Table G.16. Of the ledger's $+1.6$, $82\%$ comes from questions whose justification statements contain a perception claim about the figure, table, or diagram; the option-letter distribution ($\chi^2$ test, $p = 0.61$) and the length of the justification ($31$ vs. $30$ tokens) are unchanged.

| Category | Qwen3-VL (correct/total) | LEDGER-VLM correct per seed | LEDGER-VLM mean (SD) | $\Delta$ |
|---|---|---|---|---|
| Art & Design ($n$=120) | 102/120 = 85.0 | 105/103/105/104/104 | 86.8 (0.70) | +1.8 |
| Business ($n$=150) | 118/150 = 78.7 | 120/121/120/119/119 | 79.9 (0.56) | +1.2 |
| Science ($n$=150) | 115/150 = 76.7 | 117/118/118/119/117 | 78.5 (0.56) | +1.8 |
| Health & Medicine ($n$=150) | 120/150 = 80.0 | 121/122/121/123/122 | 81.2 (0.56) | +1.2 |
| Humanities & Soc. Sci. ($n$=120) | 103/120 = 85.8 | 104/104/103/104/105 | 86.7 (0.59) | +0.9 |
| Tech & Engineering ($n$=210) | 158/210 = 75.2 | 164/165/163/162/163 | 77.8 (0.54) | +2.6 |
| Size-weighted mean ($n$=900) | 716/900 = 79.6 | 731/733/730/731/730 | 81.2 (0.14) | +1.6 |

Table N.30: Human-study aggregates: the share of the $600$ annotator judgements per arm, three per prompt, that favour the ledger side on helpfulness, trustworthiness, and auditability, with prompt-clustered bootstrap 95% half-widths of 5–7 pp, and the relative change in mean per-prompt verification time, with annotator-stratified prompt-level bootstrap 95% half-widths of 2.4–3.2 pp. The first row is the same-answer arm and the last the no-certificate arm.

| Comparison | Help. | Trust. | Audit. | Time |
|---|---|---|---|---|
| Same answer, grounder cert. | 64.8 | 70.2 | 68.5 | $-19.7\%$ |
| vs. Qwen3-VL | 82.2 | 86.7 | 85.3 | $-38.4\%$ |
| vs. Llama-4 | 80.8 | 86.3 | 85.0 | $-37.5\%$ |
| vs. InternVL3.5 | 79.3 | 85.8 | 84.2 | $-36.2\%$ |
| vs. GPT-4o | 73.2 | 82.2 | 80.5 | $-34.7\%$ |
| vs. Post-hoc, control answer | 75.7 | 78.3 | 77.0 | $-28.1\%$ |
| vs. Claim classifier | 71.8 | 74.0 | 72.3 | $-24.6\%$ |
| No certificate, vs. Qwen3-VL | 58.3 | 57.2 | 54.8 | $+4.1\%$ |

domains under poor thresholds. Corner cases and their handling: merged claims (span-boundary preservation, adjudicated held-out set); missed detections (routed to null, never recovered); claim–atom contradictions (contradiction flag and human escalation; conflicts between evidence atoms are not modelled and remain a limitation); null-source overuse (task-specific null budgets, coverage reported with UCR); decorative certificates (counterfactual swaps make span movement a training signal and a falsification test).