VERILOOP: MULTIMODAL FACT VERIFICATION AS AN AGENTIC CLOSED-LOOP SEQUENTIAL DECISION PROCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence-based multimodal fact verification has advanced retrieval, evidence structuring, and reasoning, but a verifier must also decide what remains worth investigating and when to stop. We formulate multimodal fact verification as a closed-loop sequential decision problem and introduce VeriLoop, which maintains explicit proposition-level verification state, selects from that state what to investigate next, acquires targeted textual, visual, or cross-modal evidence, updates and replans, and stops; local proposition verification is separated from deterministic claim-level verdict derivation, so every verdict is traceable and the process itself becomes measurable. On Factify, MOCHEG, and VERITE, controlled ablations show proposition-level resolution interacts strongly with evidence quality: with graph, evidence store, and backbone fixed, it changes macro-F1 by $+9.73$ and $+1.64$ under benchmark-supplied evidence but by -9.82 under noisy retrieval, and chain-of-thought does not reproduce the effect. Reconstructing 26,417 logged interventions shows verification utility is front-loaded: first interventions dominate, most episodes end on budget, and an offline stopping-policy counterfactual shows stopping policy, verdict composition, and label granularity are coupled. Multimodal fact verification should be treated not only as reasoning over evidence, but as controlling a sequential verification process.

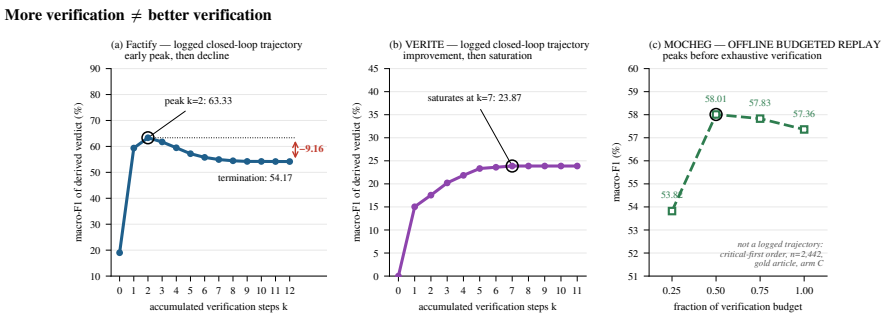

Figure 1: **More verification is not better verification.** (a) Factify and (b) VERITE: macro-F1 of the verdict derived from the accumulated state after k steps, reconstructed from the logged closed-loop trajectories; Factify peaks at k=2 and declines, VERITE saturates. (c) MOCHEG: an *offline budgeted replay* of the single-pass run (n=2,442, gold article, arm C), not a logged trajectory; macro-F1 peaks at half budget. Two different procedures, no common axis; all observational.

1 INTRODUCTION

Misinformation persists across online information ecosystems, with consequences reaching into health behaviour and resource use (Do Nascimento et al., 2022). Detection has progressed from learning patterns in content (Ma et al., 2016; Yu et al., 2017) and propagation structure (Liu et al., 2022; Li et al., 2023; Hao et al., 2024) to zero- and few-shot LLM reasoning (Chen et al., 2023;

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.

Wan et al., 2024b; Li et al., 2025a; Ma et al., 2024) and, most consequentially, to *evidence-based verification*, which judges a claim against retrieved textual or multimodal evidence rather than its surface form (Chen et al., 2024; Tahmasebi et al., 2024; Yao et al., 2023).

More evidence is not automatically more useful: redundant or mismatched material adds to the verifier's burden (Li et al., 2025a). Recent systems therefore structure evidence before judgement, via summarisation, claim decomposition, or multi-hop verification (Li et al., 2025a; Ma et al., 2024), but leave a prior question open: *after examining the current evidence, what should the verifier investigate next, and when should it stop?* For multimodal claims the question is unavoidable: one post bundles propositions with different verification requirements, and misinformation itself spans missing context, verifiable falsehood, and misleading presentation (Roozenbeek & Van der Linden, 2024). Effort spent on settled propositions is wasted. The unresolved problem is control over the verification process.

We therefore formulate multimodal fact verification as a *state-conditioned closed-loop sequential decision process*: observe the verification state, act to acquire evidence, update, replan, and decide whether to stop. LLMs supply the reasoning and tool use this requires (Wu et al., 2023; Li et al., 2025c); reliability comes from control over the whole process, not any single reasoning step. **VeriLoop** instantiates the formulation: a claim becomes a persistent *VerificationGraph* of typed propositions with explicit verification states; conditioned on that state, the loop selects the next proposition, acquires targeted textual, visual, or cross-modal evidence, verifies locally, and replans until stopping, with the verdict *derived* deterministically from the typed states. Model calls resolve propositions; the state decides the claim.

Evaluating on **Factify**, **MOCHEG**, and **VERITE** with controlled ablations and reconstructed trajectories, three findings resist the assumption that more reasoning and evidence are always better. Proposition-wise resolution is strongly moderated by evidence quality: it helps under benchmark-supplied evidence, reverses under noisy retrieval, and is not reproduced by generic chain-of-thought. Leak-free multimodal analysis shows image information contributes beyond the graph state. And verification utility is *front-loaded*: across 26,417 logged interventions, first attempts change proposition states while repeated pursuit has sharply diminishing returns; most episodes end on budget, continued pursuit coincides with degraded decisions on Factify, and a budgeted replay on MOCHEG peaks before exhaustive verification.

Contributions. (i) A formulation of multimodal fact verification as a closed-loop sequential decision process over a persistent typed proposition graph, and VeriLoop, which realises it with label-free state-conditioned planning, targeted multimodal intervention, and stopping, separating local proposition verification from deterministic claim-level derivation. (ii) Controlled evidence on three benchmarks that the value of proposition-wise resolution depends on evidence quality and is not reproduced by chain-of-thought. (iii) A characterisation of verification dynamics from logged trajectories and a frozen budgeted replay that makes evidence sufficiency and stopping central design problems.

2 Related Work

Evidence-based multimodal fact verification. Evidence-based fact-checking retrieves external evidence and classifies a claim as supported, refuted, or not-enough-information (Thorne et al., 2018; Guo et al., 2022), a paradigm set by FEVER (Thorne et al., 2018) and LIAR (Wang, 2017) and extended to multi-hop claims by EX-FEVER (Ma et al., 2024). Multimodal benchmarks carry it to image–text content: FACTIFY (Mishra et al., 2022) frames verification as five-way entailment, MOCHEG (Yao et al., 2023) pairs journalist-annotated claims with multimodal evidence and rulings, and VERITE (Papadopoulos et al., 2024) and MMFakeBench (Liu et al., 2025a) show that unimodal shortcuts and single-source assumptions overstate progress; Tang et al. (2025) add that a support-or-contradict reading misses missing context and misleading framing. Within this paradigm, transformer- and graph-based reasoners strengthen the judgement over retrieved evidence (Jiang et al., 2021; Kruengkrai et al., 2021; Zhou et al., 2019; Liu et al., 2020; Zhong et al., 2020; Xu et al., 2023; Gong et al., 2024), evidence condensation learns claim-specific multimodal summaries (Chen et al., 2024) or concept-centred reasoning support (CoReS; Li et al., 2025a), and vision-language models re-rank retrieved evidence zero-shot (Tahmasebi et al., 2024), instruction-tune for explainable out-of-context detection (Qi et al., 2024), or unify textual, visual, and cross-modal distortions in one model (Yan et al., 2025). These methods improve what evidence reaches the judgement and how well it is read,

but return a claim-level prediction and do not keep proposition-level outcomes as a state that directs further acquisition.

Structured, iterative, and agentic reasoning. A second line structures the reasoning process: fake-news-specific prompting (Huang & Sun, 2023), decomposition into verifiable sub-statements (Zhang & Gao, 2023; Wei et al., 2024), style-agnostic training (Wu et al., 2024), and knowledge-guided multi-step verification (Liu et al., 2024c); on multi-hop claims, LLMs appear more effective as planners of the process than as direct verdict predictors (Ma et al., 2024). Several systems are already iterative or tool-using: GProofT (Liu et al., 2024b) resolves sub-claims over rounds of web search with majority voting, LRQ-FACT (Beigi et al., 2025) generates targeted textual and visual questions to guide retrieval, and MMD-Agent (Liu et al., 2025a) brings tool-using LVLM agents to mixed-source misinformation; further work adds simulated social context, multi-dimensional analysis, neural-symbolic verification, sentiment-aware retrieval, and multi-agent rule optimisation (Nan et al., 2024; Wan et al., 2024a; Liu et al., 2024a; 2025b; Li et al., 2025b). Multi-step interaction alone does not secure reliability: LLMs commit early and grow less reliable as turns accumulate (Laban et al., 2026). These systems are multi-step without being closed-loop in our operational sense: the intermediate representation is not maintained as a mutable state whose observed outcomes select the next targeted intervention and decide when to stop.

Position of this work. Prior work asks how to retrieve, structure, or reason over evidence. VeriLoop asks the complementary control question: given the current verification state, which proposition remains decision-relevant, what evidence to acquire for it next, and when to stop. Its contribution is closed-loop control over persistent proposition-level verification state, with local verification separated from deterministic claim-level derivation, rather than another decomposition, retrieval, prompting, or multi-agent architecture.

3 Problem Formulation

An instance is a claim $x = (c, I)$ with text c and a possibly empty image set I, an evidence source $\mathcal{E}$ fixed by the benchmark (a corpus, a paired document, gold passages with images, or nothing), and a label space $\mathcal{Y}$. Verification is *evidence-based* when the prediction is a function of what was found in $\mathcal{E}$, which requires that the absence of evidence stay distinguishable from contradiction.

Verification state. A claim bundles propositions with different verification requirements, so we make the state explicit. A *VerificationGraph* $G = (V, E)$ has entity nodes V and proposition edges E; each edge is a tuple $e = (t_e, r_e, \kappa_e, Q_e, s_e, Z_e, a_e)$: proposition text, role $r_e \in \mathcal{R}$, criticality $\kappa_e \in [0, 1]$, search questions Q_e, state $s_e \in \mathcal{S}$, the evidence Z_e the state rests on, and attempts a_e. We group roles into assertions $\mathcal{R}_A = \{\text{ASSERTION}\}$, frame roles $\mathcal{R}_F = \{\text{PRESUPPOSITION}, \text{CONTEXT}, \text{PROVENANCE}, \text{DEPICTION}\}$, modifiers $\mathcal{R}_M = \{\text{QUANTITY}, \text{TEMPORAL}\}$, and $\{\text{ATTRIBUTION}\}$, which does not directly determine the verdict. States split into terminal $\mathcal{T} = \{\text{SUPPORTED}, \text{REFUTED}, \text{UNVERIFIABLE}\}$ and open $\mathcal{O} = \{\text{UNRESOLVED}, \text{INSUFFICIENT}\}$: REFUTED means the evidence contradicts t_e; INSUFFICIENT that evidence was examined and did not settle it; UNVERIFIABLE that no reachable source can; UNRESOLVED that it was never pursued, a fact about the budget rather than the world. We write E_A, E_F, E_M for the edges of each role class and $B(G) = \{e \in E_A : \kappa_e > 0\}$ for the *blocking* set.

Verification as a sequential decision process. Verifying one claim is an episode over states $\sigma_k = (G_k, b_k)$, the graph after k interventions and the budget spent. The legal actions are

$$\mathcal{A}(G) = \big\{(e, q, \tau) : s_e \in \mathcal{O},\ a_e < a_{\max},\ q \in Q_e,\ \tau \in \{\text{text}, \text{visual}, \text{cross-modal}\}\big\} \cup \{\text{STOP}\}. \quad (1)$$

A pursue action acquires evidence $z = \text{INTERVENE}(e, q, \tau)$ for that proposition alone, resolves it, $s'_e = \text{VERIFY}(e, z, x)$, and writes (s'_e, z) into the graph, $G_{k+1} = \text{UPDATE}(G_k, e, s'_e, z)$. The episode ends with STOP or budget exhaustion at step K, and the label is

$$\hat{y} = \text{DERIVE}(G_K) \in \{\text{SUPPORTED}, \text{REFUTED}, \text{MISLEADING}, \text{INSUFFICIENT}\}, \quad (2)$$

a deterministic function of the typed edge states that reads no raw evidence, followed by a fixed map onto $\mathcal{Y}$. Two label-free control decisions constitute the agency of the verifier: *which* proposition to pursue next, $\pi(G_k) \in \mathcal{A}(G_k)$, and *whether* to continue. Because G_k is logged, $\text{DERIVE}(G_k)$ can be evaluated at every prefix $k \le K$ of a trajectory, which is what Section 8 exploits.

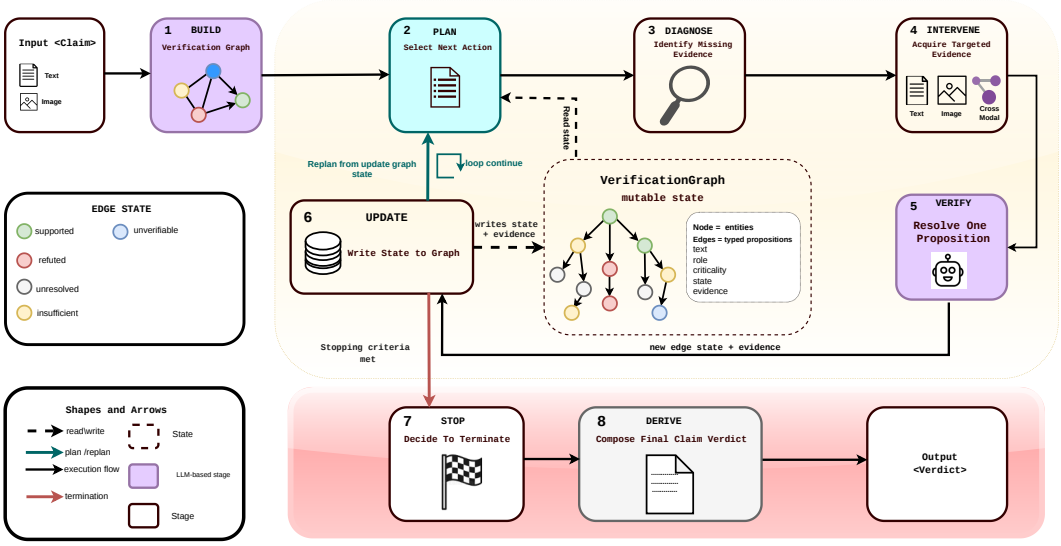

Figure 2: The VeriLoop framework. Given a multimodal claim, BUILD constructs a typed verification graph whose nodes are entities and whose edges are typed propositions carrying text, role, criticality, state, and evidence. The control loop repeatedly PLANs from the current graph, DIAGNOSEs missing evidence, INTERVENEs with targeted textual, visual, or cross-modal evidence, VERIFies one proposition, and UPDATEs the graph before replanning. When a stopping condition is met, DERIVE deterministically composes the final verdict from the typed edge states.

4 VERILOOP

VeriLoop turns multimodal verification from one-shot inference into state-conditioned closed-loop control (Figure 2, Algorithm 1). Only BUILD and VERIFY call a model; every other stage is a deterministic function of the graph. We describe the loop through four questions.

4.1 WHAT DOES VERILOOP REMEMBER?

BUILD constructs G_0 from x in one model call that decomposes the claim *by role*: what is asserted, what is presupposed, which numbers and dates it depends on, who it is attributed to, and, when an image is present, what the image shows (DEPICTION) and where it comes from (PROVENANCE). Every edge starts UNRESOLVED with one or two questions and a criticality; identifiers are content-addressed, so a claim always yields the same graph and any run replays edge by edge. Everything the loop establishes is written into G; no hidden state persists across steps.

4.2 WHAT SHOULD VERILOOP VERIFY NEXT?

PLAN reads only G. It keeps the legal actions of Eq. 1 on edges that still admit a *novel* intervention: a question not yet asked for that edge, or a wider evidence scope where one is still available; an edge with neither is retired. Each open edge is then assigned a decision-leverage category derived from the composition rule itself,

$$\lambda(e \mid G) = \begin{cases} 1 & e \in B(G),\ s_e \in \mathcal{O},\ |\{e' \in B(G) : s_{e'} \in \mathcal{O}\}| = 1, \\ 2 & e \in B(G),\ s_e \in \mathcal{O}, \\ 3 & e \in E_F \cup E_M,\ s_e \in \mathcal{O},\ \textsc{derive}(G[s_e \leftarrow \textsc{refuted}]) \neq \textsc{derive}(G), \\ 4 & \text{otherwise}, \end{cases} \tag{3}$$

where $G[s_e \leftarrow \textsc{refuted}]$ provisionally refutes e and recomposes. The planner selects the lowest category and breaks ties by criticality, then by an untried question over a widened scope of an asked

one, then by fewer attempts:

$$\pi(G) = \arg\min_{(e,q,\tau)} \big(\lambda(e \mid G),\ -\kappa_e,\ \mathbb{1}[q \text{ already asked for } e],\ a_e\big). \tag{4}$$

When the selected edge is INSUFFICIENT from an earlier attempt, DIAGNOSE recovers from the verifier's own rationale what was established, what is missing, and which channel is needed, and INTERVENE acquires evidence for that proposition only: textual (a focused lexical query over $\mathcal{E}$, or on retry the alternative question, a wider scope, or a targeted search for the missing span), visual (the claim and document images, for the roles pixels can settle or for any edge whose diagnosis asked to see), or cross-modal (both jointly). No stage consults a label, a class frequency, or a fitted weight.

4.3 What happens after evidence arrives?

VERIFY is local and model-based: one proposition and its evidence in, one edge state out. Its rubric makes absence explicit: REFUTED only when an item asserts or shows the opposite; on-topic but inconclusive evidence is INSUFFICIENT; irrelevant or empty evidence is UNVERIFIABLE, the latter without a model call; and every state must be licensed by an item the verifier itself marked. VERIFY never judges the claim and never touches G; UPDATE writes the state, confidence, evidence, and attempt count. Because the next PLAN is computed over G_{k+1}, replanning is the feedback transition itself.

4.4 How does VeriLoop finish?

STOP fires at the point where the next action would be chosen, when any of the following holds, each determined from the current episode state $\sigma_k = (G_k, b_k)$: (i) the verdict is *forced*, $\exists e \in E_A : s_e = \text{REFUTED}$, which is absorbing for DERIVE; (ii) no novel legal action remains; (iii) every edge with $\kappa_e > 0$ is terminal, or none is still pursuable; (iv) the last two pursuits changed no edge state; (v) the step cap or token budget b_k is reached. DERIVE then runs once, under the rule of record (arm C), in which only assertions block:

$$\text{DERIVE}(G) = \begin{cases} \text{REFUTED} & \exists e \in E_A :\ s_e = \text{REFUTED}, \\ \text{MISLEADING} & \text{else if } \exists e \in E_F \cup E_M :\ s_e = \text{REFUTED}, \\ \text{INSUFFICIENT} & \text{else if } \exists e \in B(G) :\ s_e \in \mathcal{O}, \\ \text{SUPPORTED} & \text{else if } B(G) \neq \varnothing \ \wedge\ \forall e \in B(G) :\ s_e = \text{SUPPORTED}, \\ \text{INSUFFICIENT} & \text{otherwise.} \end{cases} \tag{5}$$

An open frame edge therefore does not veto support, since UNRESOLVED is a fact about the budget, but a refuted one still yields MISLEADING; ATTRIBUTION does not directly determine the verdict. The rationale names the deciding edges, so every label traces to the propositions that produced it, and the verdict is mapped onto $\mathcal{Y}$, where MISLEADING has no counterpart and collapses to REFUTED by default. The separation the design protects is thus: VERIFY maps a proposition and its evidence to a state; DERIVE maps typed states to a verdict; neither does the other's job. The complete loop is given as pseudocode in Algorithm 1 (Appendix A). Section 8 examines the consequences of this stopping policy through logged trajectories and offline counterfactual analysis.

5 Experimental Setup

Benchmarks and controlled conditions. We evaluate on three complementary evidence regimes (Table 4, Appendix C): Factify provides a paired reference document and permits noisy corpus retrieval, enabling evidence-quality comparisons on the same claims; MOCHEG provides gold article- and sentence-level evidence with images; and VERITE provides no textual evidence, leaving the caption, image, and graph. Although Wikipedia evidence was acquired for VERITE, it is excluded because reverse image search can expose the benchmark's source fact-checks; all VERITE results therefore use zero retrieved evidence. Controlled ablations use one backbone (`gpt-4o-mini`, temperature 0, seed 0): `C1` predicts directly, `C2` adds only chain-of-thought, `C3` reads the graph and all proposition evidence in one flattened prompt, and `C5` resolves the same propositions individually and composes them with DERIVE. Thus, `C1`→`C2` isolates chain-of-thought and `C3`→`C5` isolates reasoning mode; `C1`→`C3` changes multiple factors and is descriptive only. The reviewer agent and partial-support rule are not evaluated (Appendix J).

Evaluation and executions. Published systems are compared under their reported splits and protocols; macro-F1 derived from reported per-class values is marked †. We report micro- and macro-F1, or accuracy and macro-F1 on VERITE under the TRUST-VL protocol (Yan et al., 2025), and retain only complete evaluation slices with zero failed calls. Closed-loop trajectories were logged for Factify (n=4,000, `claude-sonnet-5`, decision-aware controller, multimodal verification for pixel roles) and VERITE (n=714, `gpt-4o`, criticality-first controller with a DEPICTION/PROVENANCE starvation guard); MOCHEG and the ablation grid instead resolve each proposition once under the same DERIVE rule. Full loop settings and budgets are in Appendix H; the loop averages 5.5 model calls, 9.9K tokens, and $0.028 per Factify claim and 4.7 / 4.7K / $0.014 on VERITE (Table 13). Replaying every logged transition reproduces the stored verdict for every Factify and VERITE claim, supporting the trajectory analysis in Section 8; MOCHEG has no logged closed-loop trajectory, and its frozen offline budgeted replay (Appendix I) is analysed separately and never pooled with those runs.

6 MAIN RESULTS

Deterministic derivation is competitive under clean evidence, while Factify exposes a label-space mismatch. Table 1 compares VeriLoop with published systems under each benchmark's protocol, with full results in Appendix C. On three-class MOCHEG, VeriLoop-T2 achieves the highest micro-F1 among the compared systems without a learned readout; however, VeriLoop uses the gold article whereas published systems use their own retrieval, so the margin reflects evidence quality as well as method. The result therefore supports the narrower conclusion that deterministic composition of typed proposition states is competitive when evidence is clean; per-class results attribute the difference primarily to refuted precision and NEI recall (Appendix C, Section 7). Under the binary TRUST-VL protocol, zero-shot VeriLoop is statistically on par with the strongest competing result. Factify presents a different picture: published systems use the full validation split whereas VeriLoop uses its 4,000-claim closed-loop slice; CoReS-LR remains best in micro- and macro-F1, while VeriLoop leads only on the refuted class. A major readout issue is that VeriLoop's internal MISLEADING verdict has no benchmark counterpart and maps to refuted by default; remapping it to NEI improves both metrics without changing the graph, evidence, or loop, but hurts MOCHEG, identifying it as a dataset-specific adapter rather than part of the method. The supervised cross-fitted composer leads the published field, but uses a different population and is reported only as an upper reference: Factify contains a separable refuted-class artifact and even claim-text features score highly under the same cross-fitting (Appendix F), so no zero-shot or methodological claim rests on this row.

VERITE separates proposition-level visual decomposition from benchmark separability. A single CLIP ViT-L/14 image–caption cosine reaches 77.88 on the TRUST-VL protocol, above the published best (Appendix F): the benchmark is substantially separable at the embedding level. Proposition-level scoring over the same encoder reaches 79.20, the highest accuracy in Table 1c, with a significant paired gain over global similarity (Appendix E); this isolates decomposition under matched readout conditions, and the controlled conclusion is that proposition-level visual scoring improves over global image–caption similarity (Section 7). The pure configuration, with no retrieved evidence, encoder, or fitted parameter, reaches 64.15; the published systems are fine-tuned on up to 198K instruction samples or run under their own evidence regimes and serve as protocol reference points rather than controlled counterparts. Taken together, the benchmark results show that deterministic proposition-state derivation is viable with clean evidence, Factify is strongly affected by verdict–label alignment and benchmark artifacts, and supervised or CLIP-augmented results require separate interpretation from pure zero-shot verification. Accordingly, the paper's methodological claims rest on the controlled analyses of structured verification and closed-loop dynamics that follow rather than on leaderboard differences alone.

7 WHEN DOES STRUCTURED VERIFICATION HELP?

Structured verification helps when proposition-level evidence is reliable. The central controlled contrast holds the graph, evidence store, and backbone fixed and changes only whether the graph is read in one flattened prompt (`C3`) or resolved proposition by proposition and composed (`C5`). Table 2 and Figure 3 show that the effect depends strongly on evidence quality: per-proposition resolution

Table 1: Main results (%); per-class results in Appendix C. (a) MOCHEG 3-class test: T1/T2 use gold sentence/article evidence (`gpt-4o-mini`); CoReS reports 56.4 with gold multimodal evidence. (b) Factify 3-class validation: published n=7,500; VeriLoop n=4,000 (`claude-sonnet-5`, paired document); +C1 maps MISLEADING→NEI. The supervised cross-fitted readout (n=3,031) is not directly comparable[‡] (Appendix F). (c) VERITE TRUST-VL protocol (200/200): `gpt-4o`, no retrieved evidence or fitted parameters; 95% CI [62.75,66.50]. [‡] prior best; [†] derived from published per-class values; [§] readout diagnostic over an off-the-shelf encoder, not the zero-shot system (Appendix E).

(a) MOCHEG — 3-class test (n=2,442)

Method	Micro	Macro[†]
Llama3-8B	43.8	39.14
LLaVA-1.6-13B	45.4	41.82
Qwen-2.5-14B	46.3	42.01
GProofT (2024b)	46.8	43.42
MOCHEG (2023)	47.4	44.58
LVLM4FV (2024)	47.9	45.85
MSP (2024)	48.5	46.79
CoReS-MR (2025a)	51.5	47.51
CoReS-LR[‡] (2025a)	52.6	49.78
VeriLoop-T1 (sentence)	54.2	54.48
VeriLoop-T2 (article)	**57.2**	**57.36**

(b) Factify — 3-class validation

Method	F1-Ref	Micro	Macro[†]
Llama3-8B	50.8	56.0	53.29
LLaVA-1.6-13B	53.5	57.1	55.61
GProofT (2024b)	61.7	58.0	59.41
Qwen-2.5-14B	51.6	58.9	55.07
MOCHEG (2023)	62.6	59.4	51.63
LVLM4FV (2024)	61.1	61.5	59.57
MSP (2024)	66.0	62.4	62.76
CoReS-MR (2025a)	67.2	65.3	62.92
CoReS-LR[‡] (2025a)	68.1	**66.7**	**64.34**
VeriLoop Arm C	67.1	53.3	54.17
VeriLoop + C1	**87.0**	59.7	63.98
Supervised readout[‡] (cross-fitted)			
States + evidence text	—	**74.03**	**79.30**

(c) VERITE (n=486)

Method	Acc.	Macro
GPT-4o	67.50	67.57
Qwen2-VL	70.00	68.94
o1 (2024)	71.75	71.66
SNIFFER (2024)	72.50	72.02
TRUST-VL[‡] (2025)	73.75	73.61
VeriLoop (zero-shot)	64.56	64.15
raw CLIP ViT-L/14[§]	77.88	—
VeriLoop-Prop + CLIP[§]	**79.20**	**79.12**

improves macro-F1 under benchmark-supplied evidence on Factify and MOCHEG, but reverses under noisy Factify retrieval. On Factify, changing only the evidence store moves `C5` by $+11.41$ macro-F1, while `C3` is nearly invariant. Per-class analysis explains the asymmetry (Figure 4, Appendix C): with clean evidence, composition distinguishes an unsupported proposition from a false claim and restores NEI recall and refuted precision; with noisy proposition-specific evidence, it instead amplifies retrieval errors. Thus, decomposition is beneficial only insofar as the evidence assigned to each proposition is reliable; Factify's separable refuted-class artifact remains a possible untested contributor (Appendix F). The gain is also not explained by simply eliciting longer reasoning: adding chain-of-thought alone (`C1`→`C2`) is negative on both datasets (Table 2), although no paired significance test is claimed and the Factify cell is near-degenerate (Appendix C). Multimodal analyses further show a narrow but identifiable visual contribution (Appendix E): on VERITE, proposition-level image scoring significantly improves over global image–caption similarity, whereas role labels hurt and edge states add nothing over a strong encoder; on artifact-controlled Factify, image-agreement features add information beyond graph state, with the observed signal attributable to reused photographs rather than scene understanding. Importantly, reaching such visual evidence is itself a control problem: decision-relevant DEPICTION and PROVENANCE edges on VERITE were largely unattempted until a scheduling guard increased their coverage without changing the verifier.

Stopping, composition, and label granularity jointly determine the value of further verification. The Factify decline in Section 8 is not caused by a STOP–DERIVE mismatch: both use the same arm-C composition semantics (Eq. 5). Rather, even after all critical assertions are terminal, refuting a frame or modifier can change the verdict to MISLEADING, so such propositions remain decision-relevant under Eq. (3). This permits an offline stopping-policy counterfactual by truncating each frozen trajectory at the first state satisfying an alternative rule, using only edge type, criticality, and state. Table 11 and Figure 5 show that waiting for every proposition to become terminal changes no verdict, whereas stopping once critical assertions are terminal ends 2,986 of 4,000 Factify trajectories earlier and raises macro-F1 from 54.17 to 62.39. This is a per-claim population score and is distinct from the fixed-k peak of 63.33 at k=2 in Section 8. The alternative rule changes 473 verdicts, all later-produced MISLEADING outcomes concentrated in provenance, temporal, and quantity refutations; 272 are scored wrong because Factify has no MISLEADING class. Since this analysis only truncates frozen trajectories, it neither establishes a causal benefit of earlier stopping nor demonstrates generalisation to other datasets, live policies, or benchmarks that explicitly score MISLEADING. It instead shows that stopping policy, verdict composition, and benchmark label granularity are coupled and should be designed and evaluated jointly.

Table 2: Reasoning mode by evidence regime, macro-F1 (%). `gpt-4o-mini`, same graphs and evidence stores, n=4,000 (Factify) / 2,442 (MOCHEG). `C1`: direct prompt; `C2`: chain-of-thought; `C3`: one prompt over the graph and all its evidence; `C5`: per-proposition resolution with composed verdict. `C1` and `C2` read no evidence, so their Factify value does not depend on the regime. The flattened reader (`C3`) is almost invariant to the regime; the composed reader (`C5`) moves by +11.41 between retrieved and gold Factify evidence, and the contrast reverses sign.

	Factify		MOCHEG
	Retrieved	Reference	Gold article
`C1` Direct	54.25	54.25	50.23
`C2` CoT	54.12	54.12	47.02
`C3` Single	57.95	57.90	47.63
`C5` VeriLoop	48.13	**59.54**	**57.36**
Δ `C5−C3`	−9.82	+1.64	+9.73

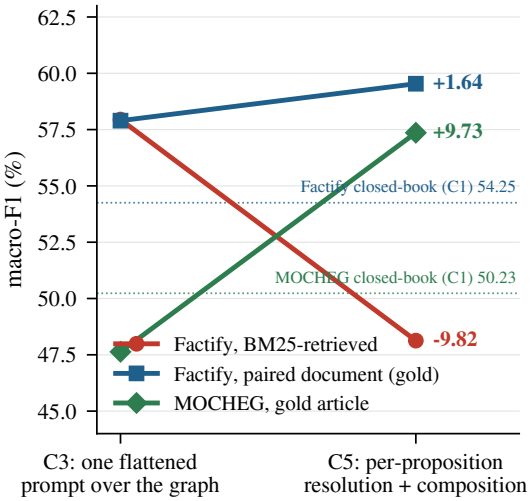

Figure 3: **Evidence quality moderates proposition-wise resolution.** Table 2 as slopes: `C3` flattened versus `C5` composed. Dotted: closed-book `C1`. Per-class shifts: Fig. 4.

8 When Does More Verification Stop Helping?

Verification utility is front-loaded and retries exhibit diminishing returns. We reconstruct every prefix of the two logged closed-loop runs; replay reproduces the stored verdict for every claim. These prefixes are states of the *same* trajectories, not independently executed or randomised budgets, and harder claims receive more steps, so later-prefix performance cannot establish a causal effect of additional verification. The Factify and VERITE runs also use different backbones, controllers, and label spaces and are never pooled; MOCHEG provides only a frozen offline budgeted replay. Within these constraints, Figure 10 shows strongly front-loaded utility. Factify peaks after two steps at 63.33 macro-F1 and declines thereafter, whereas VERITE improves monotonically but saturates at step 7 even as edge resolution continues; the latter is the strict three-class trajectory reconstruction, not the benchmark result in Table 1c. Complementarily, MOCHEG peaks at half of its offline verification budget and falls under exhaustive resolution (Figure 1(c), Appendix I). Intervention-level dynamics explain part of this pattern: first attempts usually settle previously unpursued propositions, whereas retrying an INSUFFICIENT edge reaches a terminal state in only about one third of Factify cases and rarely on VERITE (Table 3). Factify can expose new spans from its paired document after reformulation, while VERITE has no external evidence store and a retry encounters the same image. Thus the novelty constraint of Section 4.2 guarantees a *different* intervention, not an *informative* one; remaining errors are also stable under resampling, additional evidence, and a second backbone (Appendix G). The evidence therefore supports diminishing returns after early interventions, but not the causal claim that additional actions themselves reduce accuracy; the Factify decline is examined further in Section 7.

Stopping is fundamentally an evidence-sufficiency problem. Episodes typically terminate before the graph is resolved: budget exhaustion with unresolved critical edges is the dominant stopping condition, followed by the absorbing stop after a refuted assertion, while only about one quarter of Factify graphs and fewer than one tenth of VERITE graphs are fully terminal at stopping (Table 3). The different termination profiles reflect their evidence regimes: Factify contains early stops after assertion refutation but also a long tail toward the step cap, whereas VERITE, without external evidence against which to refute claims, rarely stops early and usually exhausts its budget. Together with the trajectory and replay analyses, this yields a consistent but protocol-qualified pattern: Factify improves early and then declines, VERITE improves and saturates, and the MOCHEG offline replay peaks before exhaustive verification. VeriLoop represents evidence sufficiency through its stopping rule, novelty constraint, and decision-leverage categories, but these are structural tests of the current

Table 3: **Closed-loop dynamics.** Early interventions are most productive, while extended pursuit degrades Factify performance.

	Factify (n=4,000)	**VERITE** (n=714)
Trajectory		
Logged interventions	22,092	4,325
Steps/claim, mean (median; range)	6.20 (7; 2–12)	6.97 (7; 3–11)
Best prefix Macro-F1	**63.33** (k=2)	**23.87** (k=7)
Macro-F1 at termination	54.17	23.87
Peak → termination	−9.16	**no decline**
Resolved: peak / termination	23.5% / 60.7%	54.3% / 55.3%
Intervention utility		
UNRESOLVED → terminal (first attempt)	**68.0%**	**58.0%**
INSUFFICIENT → terminal (retry)	**33.8%**	**4.0%**
Retry: no state change	**66.2%**	**96.0%**
Stopping		
Unresolved critical edges	57.3%	83.7%
Refuted ASSERTION	31.7%	8.3%
All critical edges supported	0.2%	—
Graph fully terminal	25.60%	8.68%
MOCHEG budgeted replay[§]		
Macro-F1 @ 25/50/75/100% budget	53.82 / **58.01** / 57.83 / 57.36	

[§] Offline critical-first replay (n=2,442, gold article, Arm C); not a logged closed-loop trajectory. Prefix results are observational reconstructions; Factify and VERITE runs are not pooled.

graph rather than estimates of the expected value of another intervention; a trained ordering policy and an expected-verdict-change stop were evaluated in the MOCHEG replay and not adopted (Appendix I). The results therefore identify the central control problem for an agentic verifier: not how to acquire more evidence, but whether the evidence already held is sufficient for the verdict it will report.

9 Discussion and Conclusion

We have reframed multimodal fact verification as a closed-loop sequential decision problem, realised in VeriLoop as label-free control over an explicit proposition-level verification state, with local verification separated from deterministic verdict derivation. Because the state is explicit, the verification process itself becomes an object of measurement: VERIFY maps a proposition and its evidence to a state, DERIVE maps typed states to a verdict, and a free-form verdict would admit neither replay nor counterfactual. Measured, the process shows that the reported verdict depends jointly on when the loop stops, how typed states compose, and whether the label space can express the outcome (a later frame or modifier refutation is a distinct finding about the claim that a label space without MISLEADING cannot credit), and these variables interact and must be evaluated together. Design and evaluation consequences are collected in Appendix B. The central question for a verifier is not how much evidence it can gather but what to investigate next, given the evidence it already holds, and when to stop; future systems should be designed not only to reason over evidence, but to control the verification process itself.

9.1 Limitations

Comparisons are qualified rather than like-for-like: the three benchmarks run under evidence regimes and protocols that differ from the published systems', published macro-F1 is derived from per-class values, and the Factify and VERITE runs use different backbones. All trajectory and stopping analyses are observational: prefixes come from one logged trajectory per claim and harder claims receive more steps, so no causal effect of additional actions is established; the MOCHEG replay and the stopping-policy counterfactual are frozen offline computations, and the latter does not establish that assertion-settled stopping generalises to other datasets, a live run, or a benchmark that scores MISLEADING. Decomposition is cleanly isolated only on VERITE, loop depth and top-k are unablated, and whether Factify's separable refuted class contributes to the reasoning-mode contrast is untested. A full inventory, including run-level provenance gaps, is in Appendix K.

AI USE STATEMENT

Large language models were used in a limited capacity to assist with minor editing and polishing of the manuscript, including grammar, and with the development of the interactive demonstration. All technical content, experimental design, results, analysis, and conclusions were produced, verified, and finalized by the authors.

ETHICS STATEMENT

This work evaluates an automated fact-verification method on three public benchmarks (Factify, MOCHEG, and VERITE) and involves no human subjects, no new data collection, and no personally identifying information beyond what the benchmarks already contain. Section 9 and Appendix F document benchmark artifacts that may inflate benchmark performance, and the paper's claims are bounded accordingly. Automated verifiers can produce incorrect predictions with high confidence; the deterministic and traceable verdict derivation is intended to make such errors auditable, not to replace human fact-checking.

REPRODUCIBILITY STATEMENT

Every reported result is tied to a frozen evaluation artifact. The main paper and appendix document the evaluation splits, evidence regimes, controlled conditions, verification trajectories, and provenance needed to reproduce the reported analyses. Logged transitions reproduce the stored verdicts for the closed-loop evaluations, while the MOCHEG evaluation can be recomposed from its frozen graph artifacts. Figures are generated from the frozen evaluation files without additional model calls. An anonymized interactive demo that replays the logged trajectories stage by stage is available at `https://veriloop-demo.github.io/veriloop-demo/`.

REFERENCES

Alimohammad Beigi, Bohan Jiang, Dawei Li, Zhen Tan, Pouya Shaeri, Tharindu Kumarage, Amrita Bhattacharjee, and Huan Liu. Can llms improve multimodal fact-checking by asking relevant questions? In *2025 IEEE International Conference on Big Data (BigData)*, pp. 2732–2741. IEEE, 2025.

Mengyang Chen, Lingwei Wei, Han Cao, Wei Zhou, and Songlin Hu. Can large language models understand content and propagation for misinformation detection: An empirical study. *arXiv preprint arXiv:2311.12699*, 2023.

Ting-Chih Chen, Chia-Wei Tang, and Chris Thomas. Metasumperceiver: Multimodal multi-document evidence summarization for fact-checking. *arXiv preprint arXiv:2407.13089*, 2024.

Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544, 2022.

Haisong Gong, Weizhi Xu, Shu Wu, Qiang Liu, and Liang Wang. Heterogeneous graph reasoning for fact checking over texts and tables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 100–108, 2024.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206, 2022.

Xiaorong Hao, Bo Liu, Xinyan Yang, Xiangguo Sun, Qing Meng, and Jiuxin Cao. Multi-stage dynamic disinformation detection with graph entropy guidance. *World Wide Web*, 27(2):8, 2024.

Yue Huang and Lichao Sun. Fakegpt: fake news generation, explanation and detection of large language models. *arXiv preprint arXiv:2310.05046*, 2023.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 402–410, 2021.

Canasai Kruengkrai, Junichi Yamagishi, and Xin Wang. A multi-level attention model for evidence-based fact checking. *arXiv preprint arXiv:2106.00950*, 2021.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. In *International Conference on Learning Representations*, volume 2026, pp. 54738–54778, 2026.

Fangfang Li, Zhi Liu, Junwen Duan, Xingliang Mao, Heyuan Shi, and Shichao Zhang. Exploiting conversation-branch-tweet hypergraph structure to detect misinformation on social media. *ACM Transactions on Knowledge Discovery from Data*, 18(2):1–20, 2023.

Guoyi Li, Die Hu, Haozhe Li, Qirui Tang, Xiaomeng Fu, Yulei Wu, Xiaodan Zhang, and Honglei Lyu. Zero-shot multimodal fact-checking with conceptual reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 62–71, 2025a.

Hui Li, Ante Wang, Kunquan Li, Zhihao Wang, Liang Zhang, Delai Qiu, Qingsong Liu, and Jinsong Su. A multi-agent framework with automated decision rule optimization for cross-domain misinformation detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5720–5736, 2025b.

Zhuofeng Li, Haoxiang Zhang, Seungju Han, Sheng Liu, Jianwen Xie, Yu Zhang, Yejin Choi, James Zou, and Pan Lu. In-the-flow agentic system optimization for effective planning and tool use. *arXiv preprint arXiv:2510.05592*, 2025c.

Bo Liu, Xiangguo Sun, Qing Meng, Xinyan Yang, Yang Lee, Jiuxin Cao, Junzhou Luo, and Roy Ka-Wei Lee. Nowhere to hide: Online rumor detection based on retweeting graph neural networks. *IEEE transactions on neural networks and learning systems*, 2022.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15556–15583, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.919. URL https://aclanthology.org/2024.findings-acl.919/.

Jiayu Liu, Junhao Tang, Hanwen Wang, Baixuan Xu, Haochen Shi, Weiqi Wang, and Yangqiu Song. Gprooft: A multi-dimension multi-round fact checking framework based on claim fact extraction. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 118–129, 2024b.

Xuannan Liu, Zekun Li, Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *International Conference on Learning Representations*, volume 2025, pp. 86327–86352, 2025a.

Ye Liu, Jiajun Zhu, Xukai Liu, Haoyu Tang, Yanghai Zhang, Kai Zhang, Xiaofang Zhou, and Enhong Chen. Detect, investigate, judge and determine: A knowledge-guided framework for few-shot fake news detection. *arXiv preprint arXiv:2407.08952*, 2024c.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7342–7351, 2020.

Zhiwei Liu, Kailai Yang, Qianqian Xie, Christine de Kock, Sophia Ananiadou, and Eduard Hovy. RAEmoLLM: Retrieval augmented LLMs for cross-domain misinformation detection using in-context learning based on emotional information. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16508–16523, Vienna,

Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.806. URL https://aclanthology.org/2025.acl-long.806/.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, and Shu Wu. Ex-fever: A dataset for multi-hop explainable fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9340–9353, 2024.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. Detecting rumors from microblogs with recurrent neural networks. 2016.

Shreyash Mishra, S Suryavardan, Amrit Bhaskar, Parul Chopra, Aishwarya N Reganti, Parth Patwa, Amitava Das, Tanmoy Chakraborty, Amit P Sheth, Asif Ekbal, et al. Factify: A multi-modal fact verification dataset. In *DE-FACTIFY@ AAAI*, pp. np, 2022.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 1732–1742, 2024.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C Petrantonakis. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13052–13062. IEEE, 2024.

Jon Roozenbeek and Sander Van der Linden. The psychology of misinformation, 2024. URL https://doi.org/10.1017/9781009214414.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2189–2199, 2024.

Yixuan Tang, Jincheng Wang, and Anthony Kum Hoe Tung. The missing parts: Augmenting fact verification with half truth detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33967–33984, 2025.

OpenAI Team, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. DELL: Generating reactions and explanations for LLM-based misinformation detection. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2637–2667, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.155. URL https://aclanthology.org/2024.findings-acl.155/.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*, 2024b.

William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3367–3378, 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023. URL https://arxiv.org/abs/2308.08155.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6777–6789, 2023.

Zehong Yan, Peng Qi, Wynne Hsu, and Mong-Li Lee. Trust-vl: An explainable news assistant for general multimodal misinformation detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5588–5604, 2025.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2733–2743, 2023.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. A convolutional approach for misinformation identification. In *IJCAI*, volume 2017, pp. 3901–3907, 2017.

Xuan Zhang and Wei Gao. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*, 2023.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 6170–6180, 2020.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Gear: Graph-based evidence aggregating and reasoning for fact verification. *arXiv preprint arXiv:1908.01843*, 2019.

APPENDIX OVERVIEW

The appendix supports the main paper rather than retelling it. Each section opens with the question it answers, names the protocol or control, presents the evidence, and states the bounded conclusion. Appendix A gives the full VeriLoop control loop as pseudocode (Algorithm 1); Appendix B the design and evaluation recommendations that follow from the findings; Appendix C the full leaderboards and per-class detail behind Tables 1–2; Appendix D the composition-rule contrast and the supporting detail of the stopping-policy counterfactual of Section 7; Appendix E the multimodal controls; Appendix F the benchmark-integrity diagnostics; Appendix G the error-stability and attribution analyses; Appendix H the closed-loop and graph statistics; Appendix I the MOCHEG offline budgeted replay; Appendix J the negative results and components not shipped; and Appendix K provenance and reproducibility.

A THE VERILOOP ALGORITHM

Algorithm 1 gives the complete control loop of Section 4 as pseudocode: the stages, their ordering, and which are model-based against deterministic functions of the graph.

Algorithm 1 VeriLoop: closed-loop verification of one claim x

$G \leftarrow \text{BUILD}(x)$ ▷ typed proposition graph; the only model call before the loop
while budget remains **do**
  $\mathcal{A} \leftarrow \text{LEGAL_ACTIONS}(G)$ ▷ open propositions with attempts left × questions × channels, plus STOP
  $\mathcal{A} \leftarrow \text{NOVELTY_FILTER}(\mathcal{A}, G)$ ▷ drop interventions already made on an edge; retire exhausted edges
  if $\text{STOP}(G, \mathcal{A})$ **then break** ▷ forced verdict, nothing novel, blockers settled, or no progress
  end if
  $(e, q, \tau) \leftarrow \text{PLAN}(G, \mathcal{A})$ ▷ decision-leverage category, then criticality-first tie-break
  if $s_e = \text{INSUFFICIENT}$ **then**
    $m \leftarrow \text{DIAGNOSE}(e, G)$ ▷ established / missing / channel needed, from the verifier's rationale
  else
    $m \leftarrow \varnothing$
  end if
  $z \leftarrow \text{INTERVENE}(e, q, \tau, m, x)$ ▷ textual, visual, or cross-modal evidence for e only
  $s \leftarrow \text{VERIFY}(e, z, x)$ ▷ one proposition → one edge state; model-based; does not touch G
  $G \leftarrow \text{UPDATE}(G, e, s, z)$ ▷ write state, evidence, attempts; the next PLAN is replanning
end while
$\hat{y} \leftarrow \text{DERIVE}(G)$ ▷ deterministic composition of typed edge states; reads no raw evidence
return $\hat{y},\ G$

B DESIGN AND EVALUATION RECOMMENDATIONS

Keep verification per proposition and derivation deterministic. Design the stopping rule against the verdict space the system is scored in: stop rule, composition rule, and label map form one decision, and the counterfactual of Section 7 shows how far the score moves when only the first changes. Spend retries only where new evidence can exist: novelty is necessary but not sufficient; the gating variable is whether an unexhausted source could still change the composed verdict. The error-stability results argue against ensembling a verifier whose errors are correlated across samples and backbones; no stopping policy is fitted, so every reported number is label-free. For evaluation: two of the three benchmarks are partly separable without reading their evidence and the central contrast changes sign with the evidence store alone, so a leaderboard difference is interpretable only against an artifact-driven baseline on the same split, with the evidence regime stated per row and, on Factify, the refuted class excluded.

Table 4: Evaluation benchmarks. Factify: test labels are unavailable, so the validation split is used as in the CoReS protocol; classes supported/NEI/refuted are 40/40/20. MOCHEG: balanced three-class test split. VERITE: the 714 of 1,001 captions whose image resolves, true vs. manipulated 32/68; the cross-modal protocol scores true vs. out-of-context (n=486) and all splits are claim-level because 486 photographs recur under different captions.

Dataset	Split / n	Classes	Modalities	Evidence
Factify	val / 7,500	3-class	text + image	reference document / BM25
MOCHEG	test / 2,442	3-class	text + image	gold article / gold sentence
VERITE	full / 714	2-class	text + image	none

C Extended Results: How Do the Main Comparisons Look in Full?

Full leaderboards. *Question.* Where does each system's score in Table 1 come from? *Protocol.* CoReS §5: Factify on the validation split, MOCHEG on the test split; published per-class precision, recall, and micro-F1 as reported. Tables 5 and 6 list every published system with its per-class values so that the source of each score is visible; rows under different evidence regimes are not direct comparisons. *Conclusion.* The per-class columns locate VeriLoop's differences in refuted precision and recall and NEI recall, the cells Section 7 interprets.

Table 5: Factify validation. Rule-based rows: n=4,000 closed-loop run (`claude-sonnet-5`, paired reference document). Learned composer: n=3,031 claims with retrieved evidence, 5-fold cross-fitted. Percentages. Published systems report micro-F1 only; macro-F1 is not derived here.

Method	P-sup	P-ref	P-nei	R-sup	R-ref	R-nei	micro	macro
Llama3-8B	63.6	48.5	48.8	63.5	53.4	42.6	56.0	–
LLaVA-1.6-13B	66.4	51.5	50.3	65.8	55.7	44.5	57.1	–
GProofT (Liu et al., 2024b)	68.7	53.3	51.7	66.3	73.2	46.7	58.0	–
Qwen-2.5-14B	63.4	52.8	54.3	67.3	50.5	43.5	58.9	–
MOCHEG (Yao et al., 2023)	60.4	58.8	50.5	68.8	67.0	19.3	59.4	–
LVLM4FV (Tahmasebi et al., 2024)	65.2	49.2	53.6	68.3	80.5	48.5	61.5	–
MetaSumPerceiver (Chen et al., 2024)	71.3	57.8	55.1	70.8	77.0	47.8	62.4	–
CoReS-MR (Li et al., 2025a)	72.3	57.5	55.8	66.1	80.8	49.6	65.3	–
CoReS-LR (Li et al., 2025a)	74.6	58.4	57.2	67.4	81.7	51.3	66.7	–
VeriLoop, rule-based (arm C)	54.4	51.6	54.1	46.7	95.8	38.7	53.3	54.17
+ label-space adapter `C1`	54.4	83.5	52.2	46.7	90.9	57.3	59.7	63.98
VeriLoop + learned composer (cross-fitted)	71.3	97.0	70.4	71.5	94.9	70.7	74.03	79.30

Table 6: MOCHEG test, n=2,442. VeriLoop rows: `gpt-4o-mini`, arm C, single-pass resolution, replay fidelity 100%. Percentages. Published systems report micro-F1 only; macro-F1 is not derived here.

Method	P-sup	P-ref	P-nei	R-sup	R-ref	R-nei	micro	macro
Llama3-8B	57.6	44.3	36.8	44.2	30.3	27.4	43.8	–
LLaVA-1.6-13B	59.2	45.2	40.7	48.5	33.6	28.6	45.4	–
Qwen-2.5-14B	60.5	49.1	38.3	46.5	33.9	29.5	46.3	–
GProofT (Liu et al., 2024b)	64.3	51.6	38.4	47.6	34.8	30.5	46.8	–
MOCHEG (Yao et al., 2023)	50.2	48.1	53.3	47.9	46.1	29.1	47.4	–
LVLM4FV (Tahmasebi et al., 2024)	62.9	48.2	42.8	50.1	41.8	32.6	47.9	–
MetaSumPerceiver (Chen et al., 2024)	68.8	50.3	43.8	44.9	44.2	35.1	48.5	–
CoReS-MR (Li et al., 2025a)	71.3	51.7	45.1	48.3	43.6	32.3	51.5	–
CoReS-LR (Li et al., 2025a)	73.3	53.2	46.8	51.8	46.3	33.6	52.6	–
CoReS w/ gold multimodal evidence							56.4	—
Human w/ gold multimodal evidence							70.0	—
VeriLoop, gold sentence	61.6	68.7	40.7	41.7	63.9	57.0	54.2	54.48
VeriLoop, gold article	62.4	70.1	44.0	47.1	67.6	56.6	57.2	57.36

Table 7: Cross-fitted composer on Factify validation, n=3,031 claims with retrieved evidence (%). VeriLoop's learned composer is a supervised 5-fold cross-fitted readout over graph states and evidence-side text on the stated population and is reported separately from the rule-based system; read with the benchmark-integrity analysis of Appendix F. ‡ prior best. † as in Table 1.

System	Micro-F1	Macro-F1†
VeriLoop states + evidence text	**74.03**	**79.30**
CoReS-LR‡ (Li et al., 2025a)	66.70	64.34
CoReS-MR (Li et al., 2025a)	65.30	62.92
MSP (Chen et al., 2024)	62.40	62.76

MOCHEG under the binary TRUST-VL protocol. *Question.* How does zero-shot VeriLoop compare under the multimodal-misinformation literature's protocol? *Protocol.* TRUST-VL (Yan et al., 2025) evaluates MOCHEG out of domain as balanced binary detection (200 "No" + 200 "Yes"). VeriLoop's readout is frozen ($\geq$1 refuted edge $\rightarrow$ refuted), selected on the existing run before this comparison; intervals are 10,000 balanced draws from the n=1,642 supported+refuted pool, and the full-population value without resampling is 79.66 accuracy / 79.45 macro-F1. VeriLoop consumes MOCHEG's gold article and TRUST-VL uses benchmark-provided context evidence, so the settings are comparable but not identical; the binary and three-class protocols are not comparable to each other. *Evidence.* Table 8. A secondary run with a `gpt-4o` backbone on one frozen 400-claim draw scored 78.50 [74.50, 82.50]; the larger backbone did not help. *Conclusion.* At the 400-sample resolution of the protocol, zero-shot VeriLoop and the fine-tuned model are statistically indistinguishable, and VeriLoop leads every other method in the table.

Table 8: Binary MOCHEG, TRUST-VL protocol. Published rows from Yan et al. (2025). The binary and three-class protocols are not comparable to each other.

Method	accuracy	macro-F1	setting
TRUST-VL	82.75	82.58	fine-tuned 13B LLaVA, 198K instruction samples
o1	81.50	81.38	zero-shot
VeriLoop (ours)	79.73 [76.50, 83.00]	79.49 [76.06, 82.88]	zero-shot, `gpt-4o-mini`, gold article
GPT-4o	77.00	76.81	zero-shot
LRQ-FACT (Beigi et al., 2025)	68.19	73.45	multi-LLM
MMD-Agent (Liu et al., 2025a)	54.25	43.72	multi-agent
SNIFFER (Qi et al., 2024)	53.75	50.73	specialised detector

Ablation grid: per-class detail and intervals. *Question.* What lies behind each cell of Table 2? *Control.* Every cell evaluated all claims of its slice with coverage 1.0 and zero failed calls; a first execution of the Factify gold `C5` cell lost 1,181 claims to an exhausted API quota (coverage 0.705), is excluded, and was re-run in full. *Evidence.* Table 9. Macro-F1 95% intervals: MOCHEG `C1` [48.10, 52.04], `C3` [45.71, 49.62]; Factify gold `C5` [58.10, 61.10]. Majority-class baselines: MOCHEG 33.78 accuracy / 16.84 macro-F1; Factify 40.00 accuracy. The MOCHEG gold-sentence `C5` cell scores 54.22 micro / 54.48 macro; no `C3` cell exists for that regime. *Conclusion.* The confounded `C1`$\rightarrow$`C3` contrast is $+2.93$ micro / $+3.69$ macro on Factify (retrieved) and -0.61 / -2.60 on MOCHEG; its sign reverses between datasets and it is not a component contribution.

D Composition and Stopping: What Happens Under Alternative Settings?

Composition rule: arm A versus arm C. *Question.* How much of the verdict depends on which propositions are allowed to block SUPPORTED? *Control.* Arm C (the rule of record: only ASSERTION edges block) and arm A (every edge with $\kappa_e > 0$ blocks) are two readings of the same frozen edge states, so their contrast is exact and involves no model call. *Evidence.* On the full Factify validation split (n=7,500, `gpt-4o-mini`, single pass) arm A scores 46.36 micro / 39.62 macro-F1 and arm C 48.09 / 48.21; refuted-F1 is identical (49.29) under arms A, C, and a role-aware variant R, and the

Table 9: Per-class precision and recall for the reasoning-mode and chain-of-thought contrasts (`gpt-4o-mini`, arm C, edge top-k 2).

Dataset / regime	Config	micro	macro	P-sup	P-ref	P-nei	R-sup	R-ref	R-nei
MOCHEG, gold article	`C1` direct	50.16	50.23	62.4	54.2	37.5	46.9	61.6	41.8
	`C2` chain-of-thought	47.79	47.02	68.4	53.2	36.9	28.6	62.7	52.0
	`C3` flattened	49.55	47.63	64.5	48.3	36.9	46.0	76.7	25.1
	`C5` composed	57.17	57.36	62.4	70.1	44.0	47.1	67.6	56.6
Factify, no retrieval	`C1` direct	50.70	54.25						
	`C2` chain-of-thought	54.60	54.12	52.2	85.3	47.5	13.2	75.4	86.1
Factify, retrieved	`C3` flattened	53.62	57.95	46.0	90.5	46.3	28.4	78.3	66.8
	`C5` composed	47.93	48.13	52.8	49.4	44.2	41.5	49.7	53.5
Factify, gold	`C3` flattened	53.73	57.90	45.7	90.5	46.5	27.9	78.3	67.6
	`C5` composed	56.40	59.54	53.2	72.9	50.3	42.2	84.1	56.9

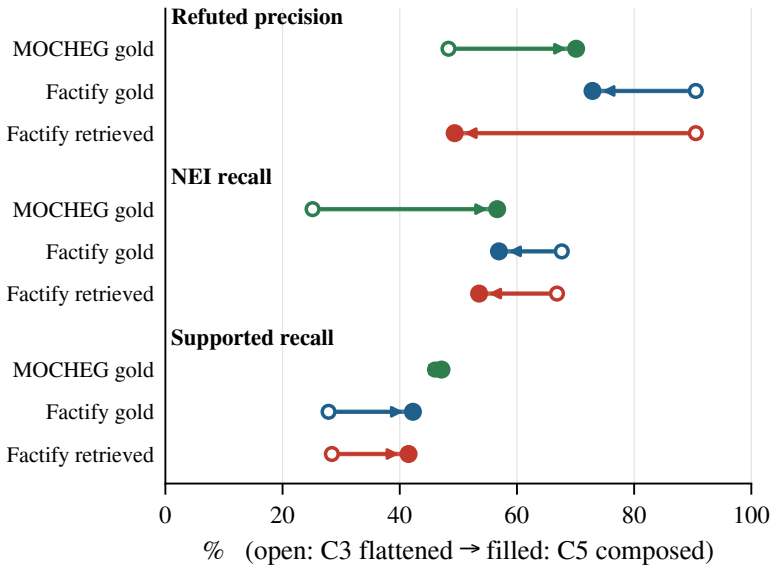

Figure 4: **Per-class shifts behind Figure 3** (open: `C3` → filled: `C5`). Clean evidence: composition restores NEI recall and refuted precision. Noisy retrieval: it surrenders the refuted precision the flattened reader keeps.

arm changes only the supported path (supported-F1 10.87 → 46.39). Re-composing the MOCHEG gold-article run offline under both arms with the canonical implementation reproduces the locked arm-C result to machine precision and gives arm A 50.2867 / 44.3426 against arm C 57.1663 / 57.3592: +6.8796 micro / +13.0166 macro for arm C, McNemar exact p=3.27 × 10^{-14} on 494 discordant claims. Every disagreement is the same move, INSUFFICIENT→SUPPORTED (529 claims; 331 right, 163 wrong); the refuted column is byte-identical. *Mechanism.* No built edge has κ_e=0 (the builder emits only 0.5 and 1.0), so arm A's blocking set is the whole graph (5.43 / 5.04 / 5.10 edges on average for Factify / MOCHEG / VERITE) against arm C's 1.84 / 1.65 / 1.49, and only 22.31% (Factify single pass), 15.60% (MOCHEG), and 8.68% (VERITE) of graphs ever become fully terminal; arm A reaches SUPPORTED on 3.65% of Factify, 3.60% of MOCHEG, and 0.14% of VERITE claims. *Conclusion.* Arm A is not a deployment configuration and affects no reported label or stopping decision. It does appear in the logs: the environment writes each STOP record's `stop_reason` by composing the graph with its own default arm, so the string "critical edges unresolved (arm A)" stands on 1,561 Factify and 546 VERITE stop records whose verdicts were composed, and whose stopping decisions were taken, under arm C. The string is trace metadata; the in-loop stopping test and DERIVE share arm-C semantics (Section 7).

Stopping-policy counterfactual: construction and detail. *Question.* What lies behind the offline stopping-policy counterfactual of Section 7? *Construction.* For each of the 4,000 Factify claims, the edge states at every step are reconstructed from the build-time states and the logged transitions; the reconstruction reproduces the stored verdict for 4,000 / 4,000 claims and the frozen termination scores (53.27 micro / 54.17 macro-F1). A candidate stopping rule is applied to each reconstructed prefix in order, the trajectory is cut at the first step that satisfies it, and the verdict is composed under arm C from the states at that step. Because the cut is a prefix of the same trajectory, every action before it is identical to the frozen run; the rules read only edge type, criticality, and state, and gold labels are used solely for scoring. No cross-claim leakage is possible through the verifier cache, which recorded zero cached steps in this run. Both rules also stop on a refuted assertion. The *strict* rule stops once every assertion, frame, and modifier edge is terminal; the *assertion-settled* rule once every critical assertion is terminal. *Evidence.* Table 10. Under the strict rule 1,132 claims stop earlier than the frozen run (mean step 5.90 against 6.20) and no verdict changes. Under the assertion-settled rule the stop step is 1 for 1,605 claims, 2 for 880, 3 for 327, and at most 12; seven claims have no critical assertion and stop at step 0 with INSUFFICIENT. Micro-F1 under that rule is 59.40. All 473 changed verdicts are MISLEADING at the frozen termination: 350 become SUPPORTED (181 wrong→correct, 146 wrong→wrong, 23 correct→wrong) and 123 become INSUFFICIENT (91, 28, 4). *Conclusion.* The frozen loop never continued past a fully determined verdict; what the later steps did was refute frame and modifier propositions after the assertions had settled. The counterfactual is a truncation of frozen trajectories, not a live run, and its population score under per-claim stopping is not a fixed-k prefix point; it does not establish that assertion-settled stopping generalises to other datasets, to live execution, or to a benchmark that scores MISLEADING.

Table 10: Stopping-policy counterfactual on the frozen Factify trajectories (n=4,000, arm C): supporting detail for Section 7. Left: refutations logged after the assertion-settled stop step, by proposition role (all claims / claims whose verdict changes). Right: the 473 changed verdicts by target verdict and scoring outcome. Offline truncation of logged trajectories; no live run.

Role	all	changed
provenance	212	212
temporal	166	166
quantity	110	110
depiction	88	88
attribution	31	13
context	19	19
presupposition	13	13

MISLEADING →	W→C	W→W	C→W
SUPPORTED	181	146	23
INSUFFICIENT	91	28	4

Stop step under the assertion-settled rule

0: 7; 1: 1,605; 2: 880; 3: 327; 4: 193; 5: 214; 6: 179; 7: 143; 8: 162; 9: 136; 10: 90; 11: 38; 12: 26

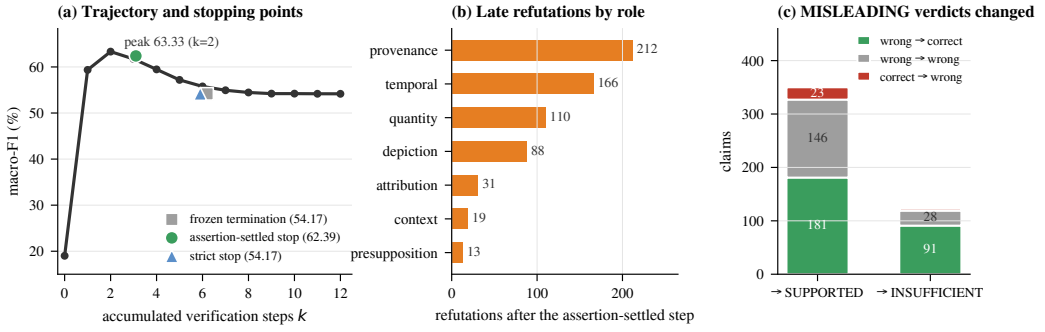

Figure 5: Stopping-policy counterfactual on the frozen Factify trajectories. (a) Macro-F1 across accumulated steps, with the mean stopping step of each rule marked; the marker heights are population scores under per-claim stopping, not points on the fixed-prefix curve. (b) Refutations logged after the assertion-settled step, by proposition role. (c) The 473 verdicts that stopping earlier changes, split by scoring outcome. Offline truncation of logged trajectories; no live run.

Table 11: Stopping-policy counterfactual on the frozen Factify trajectories (n=4,000, arm C). Per-claim stopping once the assertions are settled raises the population macro-F1 to near the fixed-prefix peak of 63.33 (k=2); the rows are population scores under per-claim stopping, not fixed-k prefix points.

Stop rule	Micro	Macro	Stop earlier	Mean step	Changed	W→C	C→W
Frozen run	53.27	54.17	—	6.20	—	—	—
Strict: every edge terminal	53.27	54.17	1,132	5.90	0	0	0
Assertion-settled	**59.40**	**62.39**	2,986	3.10	473	**272**	27

Both rules also stop on a refuted assertion and read only edge type, criticality and state. Stop earlier: claims whose counterfactual stop precedes the logged termination. Changed: verdicts that differ at the counterfactual stop; W→C / C→W: wrong-to-correct and correct-to-wrong under gold, used for scoring only. Counterfactual truncation of logged trajectories, not a live run.

E Multimodal Evidence: What Does the Image Channel Contribute?

VERITE probe. *Question.* What does each VeriLoop layer add over a raw CLIP image–caption similarity? *Protocol.* Readout diagnostics, distinct from the pure zero-shot system result of Table 1c: paired over 30 identical claim-level 70/30 splits (ViT-L/14, n=486, fixed configurations, no selection on test folds). *Evidence.* Table 12 and Figure 6. Decomposition into untyped propositions is the only significant positive effect; role typing is significantly negative; graph edge states are null against ViT-L/14. The contribution of edge states depends on the encoder: with ViT-B/32, typed propositions × CLIP score 67.65 and adding edge states 72.23 ($+4.58$); with ViT-L/14, 77.42 and 77.26 (-0.16). The raw-CLIP baseline itself (77.88) already exceeds the published best (TRUST-VL 73.75) by 4.13, so the paired deltas are measured against that bar. Backbone ablation of pure zero-shot VeriLoop: `gpt-4o` 64.56 / 64.15 accuracy / macro-F1, `qwen3-vl-32b-instruct` 61.86 / 60.99. *Conclusion.* Structured state compensates for a weak perceptual front-end and is redundant against a strong one; decomposition is the one layer that helps against either.

Table 12: VERITE cross-modal, accuracy over 30 identical splits with 95% intervals on the mean and paired differences to the raw-CLIP baseline.

Configuration	acc.	95% CI	Δ vs. raw CLIP	95% CI (t)
raw CLIP(image, caption)	77.88	[77.08, 78.67]	—	
CLIP × VeriLoop propositions, untyped	79.20	[78.35, 80.05]	$+\mathbf{1.32}$	[+0.73, +1.92] (4.35)
CLIP × propositions, typed roles	77.42	[76.41, 78.43]	-0.46	[−1.29, +0.37] (−1.08)
role typing alone (typed − untyped)			$-\mathbf{1.78}$	[−2.55, −1.01] (−4.51)
CLIP × typed propositions + graph edge states	77.26	[76.16, 78.36]	-0.62	[−1.81, +0.58] (−1.01)
graph edge states only, no CLIP	67.05	[65.78, 68.33]	-10.82	[−12.47, −9.17] (−12.86)

Factify: retracted table and leak-free analysis. *Question.* Does image information add to the graph state once the benchmark's image-presence artifact is removed? *Control.* A full three-class multimodal feature ablation (graph states .6410 → states+text+all image features .7384 micro-F1) was retracted after a leakage decomposition showed that image presence and size alone deliver .7146, 7.4 of the 9.7 points, through the artifact of Appendix F; a -1 encoding for absent images carried the same leak into the agreement features. The leak-free analysis restricts to the 2,609 claims with both images and to supported versus NEI, where the artifact cannot act (the agreement block's minimum is exactly zero). *Evidence.* Image-agreement features raise a graph-state readout from 59.83 to 65.89 micro-F1 ($+6.06$), states plus hand text reach 64.66, and states, text, and image agreement together reach 67.65, above a claim-text bag of words (66.69), which states plus image agreement alone (65.89) do not (Figure 7). Single-feature AUCs: size ratio .656, average-hash distance .564, histogram L_1 .562, aspect-ratio difference .560, brightness difference .544, difference-hash .539, near-duplicate .503. *Conclusion.* The mechanism is a reused photograph, identical image area occurring on 21.8% of supported and 4.3% of NEI claims, not scene understanding.

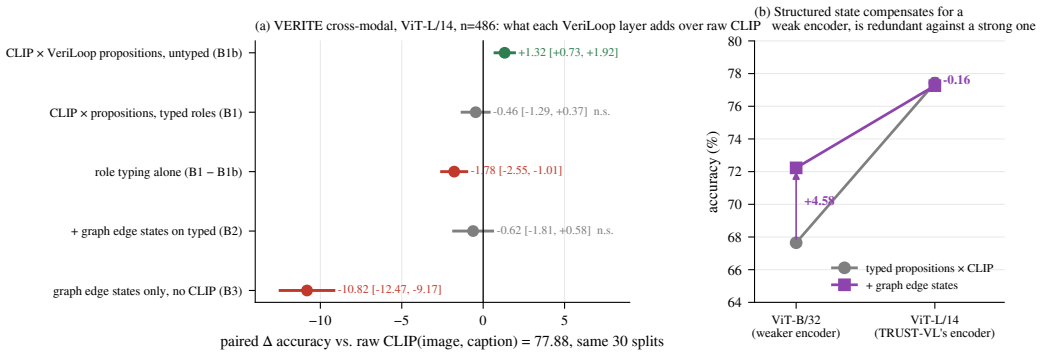

Figure 6: **VERITE: what each VeriLoop layer adds over a raw CLIP similarity.** (a) Paired accuracy differences over raw CLIP(image, caption) on the same 30 claim-level splits (ViT-L/14, n=486) with 95% intervals: decomposition into untyped propositions is significantly positive, role typing is significantly negative, and graph edge states are null. (b) Encoder dependence: edge states help against ViT-B/32 and are redundant against ViT-L/14.

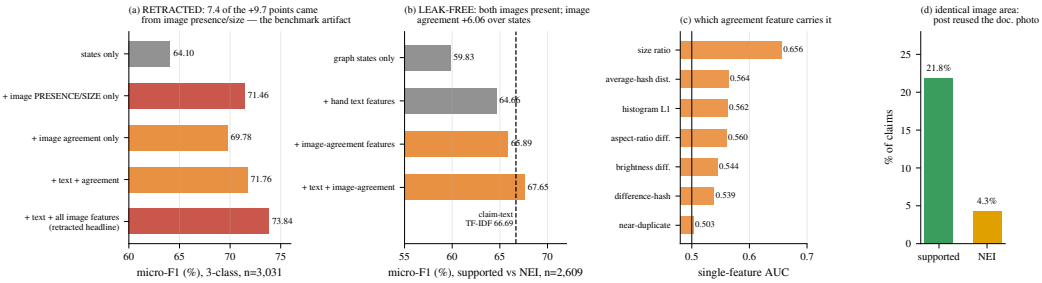

Figure 7: **Factify image channel: what was a leak and what survives.** (a) The retracted full-feature table: image presence/size alone delivers most of the gain through the benchmark artifact. (b) Leak-free cohort (both images present, supported vs. NEI, n=2,609): image-agreement features add to graph states, and states, text, and agreement together edge past a claim-text bag of words, which states plus agreement alone do not. (c) Single-feature AUCs. (d) Mechanism: the post reused the document's photograph far more often on supported than on NEI claims.

VERITE coverage guard. *Question.* Does the loop reach the visual propositions at all? *Evidence.* Replaying frozen VERITE traces showed 46 of 49 DEPICTION and PROVENANCE edges never attempted across 23 claims, all at criticality 1.0: 47.8% because a refuted assertion had made the verdict forced before the frame edges were reached, 52.2% because the no-progress condition fired inside the tie-break after the assertions were exhausted. *Control.* A scheduling-only guard (verifier, thresholds, composition, and graph builder untouched) intervenes only when the controller would stop while an eligible unattempted decision-relevant visual edge remains. It removed the starvation ($46 \rightarrow 0$), newly resolved 22 edges with 0 regressions, and raised the strict three-class result by $+4.09$ micro / $+6.17$ macro-F1. *Conclusion.* On VERITE the five non-visual roles terminate with 0% supported and 0% refuted, all UNVERIFIABLE, while only the three roles that receive pixels (ASSERTION, DEPICTION, PROVENANCE) obtain substantive verdicts: the clearest evidence in the artifacts that absence of evidence is not converted into refutation.

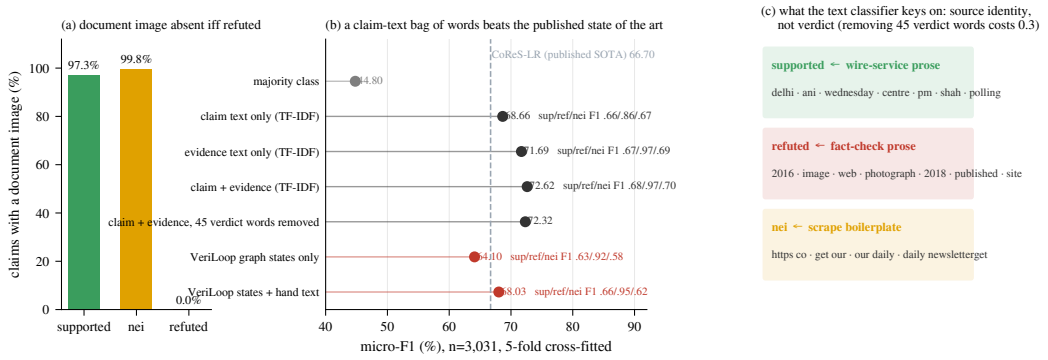

Figure 8: **Factify benchmark integrity.** (a) A document image is present for nearly every supported and NEI claim and for no refuted claim. (b) Under 5-fold cross-fitting a claim-text bag of words already exceeds the published state of the art, and evidence text alone reaches perfect refuted precision. (c) The text classifier keys on source identity (wire-service, fact-check, and scrape vocabulary), not on verdict words: removing 45 verdict words costs 0.3 points.

F Benchmark Integrity: How Much Is Separable Without Reading Evidence?

Factify. *Question.* How much of Factify can be scored without reading the evidence? *Evidence.* On the validation split the document image is present for 97.3% of supported and 99.8% of NEI claims and for 0.0% of refuted claims: the image is absent if and only if the claim is refuted (Figure 8a). Recovery of the missing images through the original URLs returned 18 of 745; all 533 Snopes URLs return HTTP 403. Under 5-fold cross-fitting on the 3,031 claims with retrieved evidence, a claim-text-only TF-IDF classifier, which is not a proposed method, scores 68.66 micro-F1, above the published reference point on this split (66.70, CoReS-LR); evidence text only scores 71.69 (refuted precision 1.00), claim plus evidence 72.62, and the same with 45 verdict words removed 72.32, a 0.3-point cost, so the signal is source identification rather than verdict leakage (top learned terms: wire-service tokens for supported, fact-check prose for refuted, scrape boilerplate for NEI). *Conclusion.* Any multimodal model scores the refuted class perfectly without reading content, which is why the learned composer is reported separately (Table 7) and why the multimodal analysis of Appendix E is restricted to claims with both images. The Factify text-only stratum (n=727) is 95% refuted by construction and is not a comparable population.

VERITE. *Question.* How much of the cross-modal split is separable by embedding similarity alone, and what is reachable from pixels? *Evidence.* A single cosine similarity between image and caption (CLIP ViT-L/14, the encoder TRUST-VL uses) scores 77.88 accuracy on the cross-modal split, above the published best (73.75). The miscaptioned versus out-of-context distinction is at chance for VeriLoop (63.0% against a 63.0% majority; four independent tests agree), because out-of-context is defined by external provenance that is unreachable from pixels. Full-image reverse search surfaces the answer key, 15.3% of verified full matches being `snopes.com`, the source VERITE was built from; on one judge Wikipedia passages reached 24.1% target-sufficiency against 1.1% for web detection. *Conclusion.* External evidence was acquired for all 714 claims but deliberately not evaluated, and every VERITE result in the paper is a zero-retrieved-evidence result.

G Error Stability: Are the Remaining Errors Noise or Verifier Error?

Errors are stable. *Question.* Would resampling, more evidence, or a second model remove the remaining errors? *Evidence.* On Factify, assertion errors are unanimous under three-sample self-consistency (92.7%, n=55); 26 of 27 persist when the verifier is given the full evidence pool and 20 of 26 under a disjoint evidence view; and a second backbone (Qwen) makes the same mistakes, both

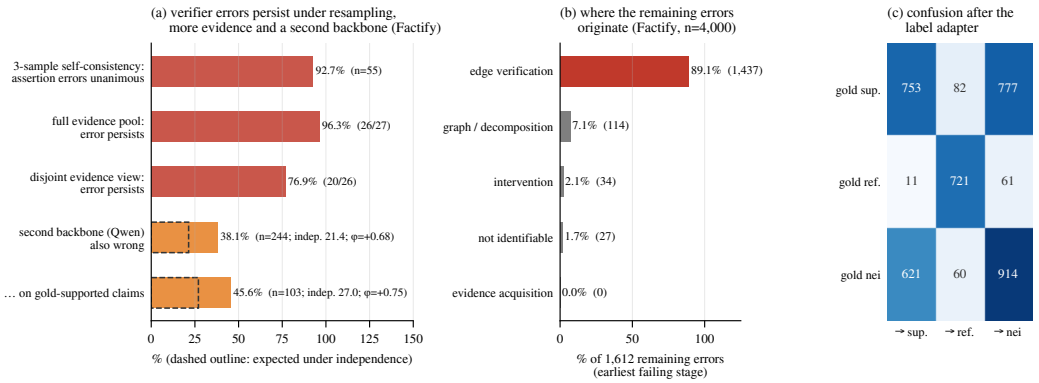

Figure 9: **Verifier errors are stable and originate at verification.** (a) On Factify, assertion errors are unanimous under three-sample self-consistency, persist with the full evidence pool and a disjoint evidence view, and are shared by a second backbone far above the rate expected under independence (dashed). (b) Earliest failing stage of the remaining errors after the label adapter: edge verification accounts for nearly all, evidence acquisition for none. (c) The remaining confusion lies on the supported–NEI axis.

wrong on 38.1% of 244 paired claims against 21.4% expected under independence ($\phi = +0.676$), and on gold-supported claims 45.6% against 27.0% ($\phi = +0.747$), where even a perfect two-model oracle would reach only 54.4% (Figure 9a). Self-consistency voting with three samples yields +2.00 micro-F1 (p=0.40) with Claude and +1.64 (p=0.13) with Qwen, neither significant. *Conclusion.* The errors are correlated, not independent; ensembling and more evidence do not remove them, which is the basis of the anti-ensembling implication in Section 9.

Where errors originate. *Question.* At which stage of the loop does an error first become observable? *Evidence.* After the label-space adapter, 1,612 errors remain on the n=4,000 Factify run. Attributing each to the earliest observable failing stage gives edge verification 89.1% (1,437), graph decomposition 7.1% (114), intervention 2.1% (34), not identifiable from the trace 1.7% (27), and evidence acquisition 0.0% (Figure 9b). The remaining confusion lies on the supported–NEI axis: supported→NEI (777) and NEI→supported (621) together carry 86.7% of all remaining error (Figure 9c). *Conclusion.* Evidence acquisition is never the earliest failure; the verifier's judgement of a proposition against evidence already present is.

H Closed-Loop and Graph Statistics: How Do Trajectories Behave in Detail?

Loop settings for the executions of Section 5: at most 12 steps and one token budget per claim, at most two attempts per edge (three under the decision-aware controller), and per-edge top-k of 2 for the ablation grid and MOCHEG and 4 for the closed-loop runs.

Controller statistics. *Question.* Is criticality steering the budget, and where does effort go? *Evidence.* Table 13. The decision-relevant roles, ASSERTION, DEPICTION, and PROVENANCE, are the hardest to move on both datasets and receive the most attempts per edge (Factify 1.184, 1.058, 1.089 against 0.456 for PRESUPPOSITION), and assertions are over-pursued relative to their structural share (39.44% of pursuits vs. 33.90% of edges on Factify; 32.67% vs. 29.33% on VERITE). Factify's termination depth is bimodal, a 17.2% spike at the minimum depth where the decision-aware controller stops after one decisive refutation and a second mode at 6–7 steps, whereas VERITE, without that early stop, is unimodal at 5–7. *Conclusion.* Criticality visibly steers the budget toward the roles that decide the verdict.

Table 13: Efficiency, termination, and per-role effectiveness of the two closed-loop runs (logged values).

	Factify (n=4,000)	VERITE (n=714)
Controller	decision-aware, criticality-first tie-break	criticality-first + coverage guard
Verifier backbone	`claude-sonnet-5`	`gpt-4o`
Evidence store	paired reference document	none
Pursuits per claim, mean / median / sd	5.52 / 6 / 2.77	6.06 / 6 / 1.42
Model calls, tokens, cost per claim (mean)	5.52, 9,856, \$0.0279	4.73, 4,743, \$0.0142
Edges never attempted / retried ($\geq$2)	22.83% / 21.31%	6.54% / 25.43%
Attempt histogram (1st / 2nd / 3rd)	16,780 / 4,634 / 678	3,400 / 925 / —
Actions: primary / alternative / expand scope	75.96 / 8.84 / 15.21%	78.61 / 21.39 / 0%
State changed on attempt 1 / 2 / 3	100 / 34.9 / 25.8%	100 / 4.0 / —
Share of interventions that changed the edge's state, by role		
presupposition / context / attribution	96.1 / 91.5 / 89.3%	100 / 100 / 100%
temporal / quantity	87.1 / 85.3%	100 / 100%
assertion / depiction / provenance	83.8 / 82.5 / 80.5%	75.0 / 72.9 / 72.7%
Termination depth, % of claims stopping after exactly k steps		
$k = 2$ / 3 / 4 / 5 / 6	17.20 / 3.25 / 4.50 / 10.72 / 14.20	— / 0.28 / 6.02 / 8.40 / 25.07
$k = 7$ / 8 / 9 / 10 / 11 / 12	16.12 / 15.10 / 9.68 / 5.08 / 1.90 / 2.25	20.31 / 27.03 / 8.12 / 4.20 / 0.56 / —

Stop records. The `stop_reason` written on a STOP record is the composition rule the graph composed to at that moment under the environment's default arm (Appendix D), not which of the controller's label-free conditions fired; the latter are not logged. Factify has 2,723 stop records (1,277 episodes ended on the environment's limits without one) and VERITE 652 (62). A post-hoc diagnostic that uses gold labels the controller never sees: Factify claims scored correctly received 5.42 steps on average (n=2,131) and incorrect ones 7.10 (n=1,869); VERITE 5.94 (n=173) vs. 7.30 (n=541). This is the expected signature of a controller that works longest on inputs it cannot settle and is not evidence about any runtime stopping policy.

What the VerificationGraph encodes. *Question.* What do the built graphs look like, and does their state carry label signal without seeing a label? *Evidence.* Graphs are built once per dataset and reused by every run: Factify 87,256 nodes and 40,719 edges over 7,500 graphs; MOCHEG 25,038 / 12,296 over 2,442; VERITE 7,063 / 3,638 over 714. Every dataset has a median of five edges and ten to eleven nodes; what differs is variance (edge-count sd 1.05 / 0.85 / 0.42). ASSERTION is 29–34% of edges and present in 99.5–100% of graphs; MOCHEG is the only dataset where PRESUPPOSITION is a major role (10.23% of edges, prevalence 51.19%); VERITE has PROVENANCE in 100% and DEPICTION in 96.5% of graphs. Every edge carries at least one question (mean 1.42 / 1.57 / 1.66), so the action space is never structurally empty. Criticality takes only the values 0.5 and 1.0 (never 0) and ranks roles consistently on all three datasets (ASSERTION $\approx$ PROVENANCE $>$ DEPICTION $>$ TEMPORAL $\approx$ ATTRIBUTION $>$ PRESUPPOSITION $>$ CONTEXT). Final edge states in the Factify closed-loop run: unresolved 22.83%, supported 37.29%, refuted 7.75%, insufficient 16.46%, unverifiable 15.67%; in VERITE: 6.54 / 12.34 / 4.89 / 38.21 / 38.02%. *Conclusion.* Every role refutes more often on gold-refuted than on gold-supported claims on every dataset without the builder or verifier seeing a label; on VERITE the visual roles carry the signal (DEPICTION refutes on 14.5% of manipulated vs. 2.3% of true captions, PROVENANCE 10.1% vs. 0.9%).

I MOCHEG UNDER A VERIFICATION BUDGET (OFFLINE BUDGETED REPLAY)

Question. How does MOCHEG performance depend on the verification budget, given that it has no logged closed-loop trajectory (Section 8)? *Construction.* The frozen policy-layer evaluation replays the single-pass MOCHEG run (test, n=2,442, gold article, arm C, full state coverage) *offline* under a verification budget: each policy orders the propositions of a claim and is allowed a fraction of the full budget, and the verdict is composed from the frozen edge states reached within it. This is an offline budgeted replay; it is not a logged closed-loop trajectory, and the two are never pooled. At full

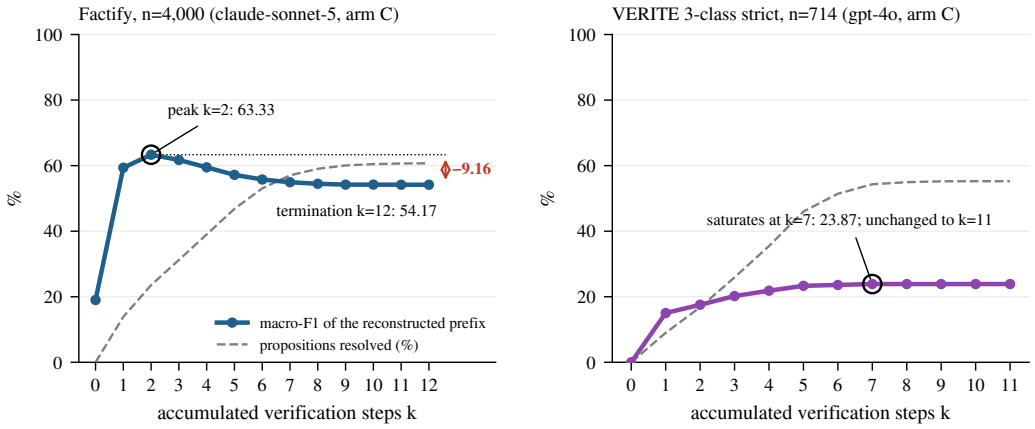

Figure 10: **Verification utility is front-loaded.** Macro-F1 of the verdict derived from the reconstructed graph after k accumulated steps (solid) and share of propositions resolved (dashed), for the Factify and VERITE closed-loop runs. On Factify the accumulated state scores best after two steps and declines thereafter to termination; on VERITE the curve rises monotonically and saturates at k=7. The VERITE run scores below the majority baseline in its strict three-class reading, so its panel describes loop dynamics, not competitive performance. Prefixes are reconstructed states of one trajectory per claim; read as performance across accumulated steps, not as an evidence-scaling law. MOCHEG has no logged closed-loop trajectory; its complementary offline budgeted replay is reported in Figure 1(c) and Appendix I.

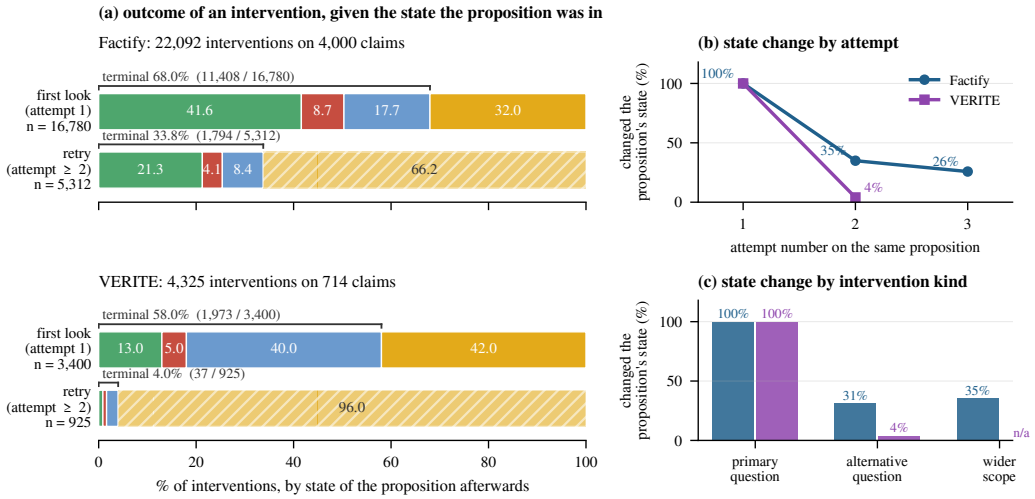

Figure 11: **Why repeated acquisition has diminishing returns.** (a) Outcome of an intervention given the state the proposition was in when pursued: a first look on an UNRESOLVED proposition (attempt 1) versus a retry on a proposition already left INSUFFICIENT (attempt $\geq$ 2), as the distribution of the state the verifier returned; brackets mark the share reaching a terminal state and hatching marks retries that left the state unchanged. (b) Probability that an intervention changes the proposition's state, by attempt number. (c) The same by intervention kind: the primary question, the edge's alternative question, or a wider evidence scope (unused on VERITE, whose store is empty). Values are read directly from the pursuit logs of the two closed-loop runs, not reconstructed.

budget every non-stopping policy converges to the locked MOCHEG result (57.36 macro-F1 / 57.17 micro-F1), which the replay harness reproduces exactly.

Table 14: MOCHEG offline budgeted replay — not a logged closed-loop trajectory. Macro-F1 by fraction of the verification budget, arm C. Values reproduced verbatim from the frozen policy-layer evaluation.

policy	0.25×B	0.50×B	0.75×B	1.00×B
critical-first (shipped heuristic)	53.82	58.01	57.83	57.36
all-edges / exhaustive order	52.78	57.59	57.10	57.36
trained π	52.87	57.82	57.94	57.36
EVC-guided, θ=0.02	51.67	57.08	57.29	57.68
EVC-guided, θ=0.05	51.53	57.05	57.32	57.71
EVC-guided, θ=0.10	51.28	56.76	56.87	56.85
random order	33.21	44.44	50.75	57.36

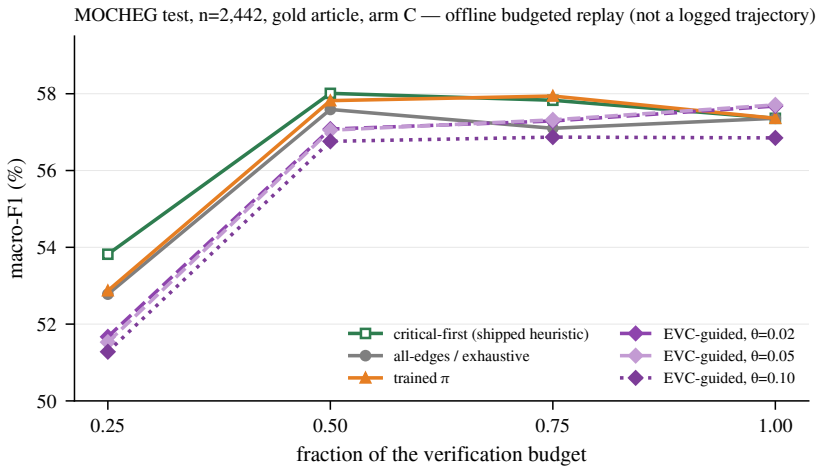

Figure 12: **MOCHEG under a verification budget — offline budgeted replay; not a logged closed-loop trajectory.** Macro-F1 of the single-pass MOCHEG run (test, n=2,442, gold article, arm C) when each policy is allowed a fraction of the full verification budget, replayed offline over the frozen edge states. Under the shipped critical-first order macro-F1 peaks at half the budget and is lower under exhaustive resolution; the decision-relevance stop (EVC-guided) reaches its best value at full budget while pursuing fewer propositions. Values reproduced verbatim from the frozen policy-layer evaluation; the full-budget point reproduces the locked MOCHEG result.

Evidence. Table 14 and Figure 12. Under the shipped critical-first order macro-F1 peaks at 58.01 at half of the budget and is 57.36 under exhaustive resolution. The decision-relevance stop (EVC-guided, θ=0.05) reaches 57.71 at full budget while pursuing 2.984 of 5.035 propositions per claim and 5.10M against 8.61M nominal tokens, which the frozen evaluation records as a real efficiency result even though the EVC policy was not adopted (it fails non-inferiority at the tighter budgets). The frozen evaluation also reports the composition arms on this run (arm A 44.34, arm C 57.36, arm R 52.62 macro-F1), consistent with Appendix D. *Conclusion.* MOCHEG performance peaks before exhaustive verification rather than improving monotonically with budget. No trajectory, transition, or retry statistic is available for MOCHEG, and none is inferred.

J Negative Results and Components Not Shipped

Figure 13 summarises every documented mechanism by dataset. The entries below constrain what may be claimed; none is part of the evaluated system unless stated.

Label-space adapter `C1`. Of 4,000 Factify claims, 609 compose to MISLEADING; all 609 mapped to refuted by default, and their gold labels are NEI 296, supported 274, refuted 39. Remapping MISLEADING to NEI changes exactly those 609 predictions (+6.43 micro, CI $[+5.55, +7.30]$;

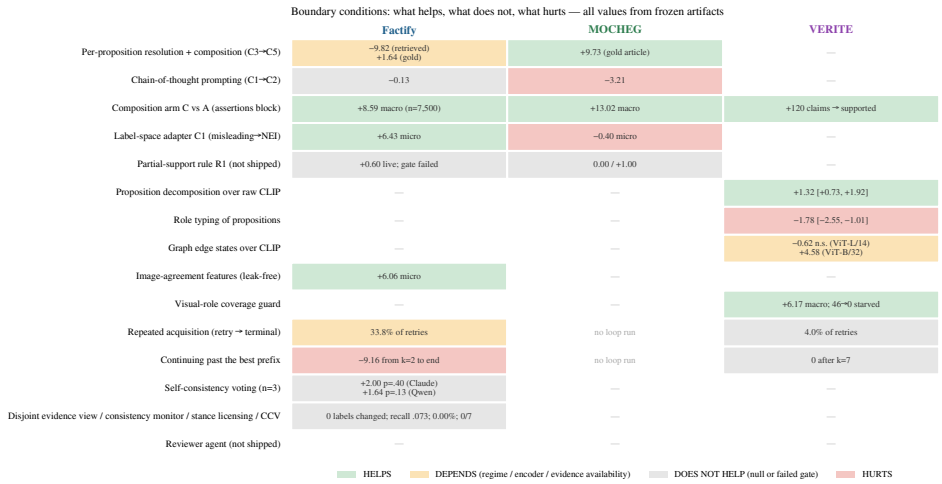

Figure 13: **Boundary conditions.** Each documented mechanism by dataset, coloured by whether it helps, depends on the regime, is null, or hurts, with the frozen value in the cell. Almost nothing helps unconditionally: composition and decomposition help where evidence is clean or absent by design; chain-of-thought, role typing, and continued pursuit hurt; retries, self-consistency, and the auxiliary monitors are null.

296 wrong→right, 39 right→wrong) and touches no other claim, so its offline estimate equals its prospective value by construction (100% replay fidelity). On MOCHEG it costs 0.40 micro under both evidence regimes. It is a readout correction of a mis-specified mapping, not a component.

Partial-support rule `R1` (not shipped). Estimated offline at $+1.78$ micro on Factify; live and prospective on $n{=}501$ paired claims it measured $+0.60$ and failed its attribution gate (64%). Unlike `C1`, `R1` changes the composition that the planner and stopping test read, so going live changed when the loop stopped; the offline estimate overstated it about three-fold. Whether a mechanism feeds back into planning is therefore the property that decides whether its offline estimate transfers; the stopping-policy counterfactual of Appendix D is a pure truncation and does not have this exposure, which is why it is reported as offline rather than as a live result.

Reviewer agent and L3 reconsideration (not shipped). The evaluated configuration runs with the reviewer disabled: in development ablations it provided no measurable benefit and mediated negotiation was subtractive, so both are excluded from the evaluated system. The single diagnostic-conditioned reconsideration (L3) was evaluated and did not meet its adoption criterion; it is not adopted.

Null and negative auxiliaries. A disjoint-evidence view changed 4 edges and 0 labels; a consistency monitor had precision .80 but recall .073; stance licensing changed 0.00% of edges; condition-level verification (CCV) rescued 0 of 7; an open backbone variant (L1) was -4.90. Role typing on VERITE is significantly negative (Appendix E); edge states over a strong encoder are null. The composition arm and the loop depth were not ablated at full scale beyond what Appendix D reports; `edge_top_k` and `planner_top_k` are unablated (fixed at 2 or 4, and 3).

K Provenance and Reproducibility

All numbers are read from committed artifacts: the locked result documents, the verbatim evaluation reports of every ablation cell, the frozen per-claim records of the two closed-loop runs and the MOCHEG system run, and the frozen graph artifacts. Every figure is regenerated by a script from those files with no model call; the stopping-policy counterfactual of Appendix D is likewise a script over the frozen Factify records that first reproduces the locked termination and prefix values. Coverage gates: a cell backs a table only if it evaluated every claim of its slice with zero failed

calls; the one truncated execution (Factify gold C5, coverage 0.705 after an exhausted API quota) is preserved for audit and excluded from every result.

The MOCHEG system run executed at a clean-tree commit. The Factify and VERITE closed-loop runs and the ablation grid ran against a dirty working tree, and their manifests do not tie them to a commit; the MOCHEG C1/C3 cells reproduce an earlier execution to four decimals, establishing determinism but not provenance. The evidence store consumed by the MOCHEG C5 run is not recorded in its manifest; directory timing and naming are consistent with the gold-article store the C3 cell was given. The VERITE runner input file is not archived; its 714-claim cohort was reconstructed independently from the released CSV and reproduces the graph identifiers exactly. Replaying logged transitions reproduces the stored verdict for 4,000 / 4,000 Factify and 714 / 714 VERITE claims, and re-composing the MOCHEG run with the canonical implementation reproduces its locked scores to machine precision.

What cannot be reconstructed: the modality of each intervention (images reach the verifier inline and are never persisted as evidence records; the logged tool is corpus on 100% of pursuits); which of the controller's stopping conditions fired (only the composed rule, written under the environment's default arm, is on the stop record); the candidate action sets at each decision (only the selected action is persisted); and evidence stance (never populated). The frozen implementation was not modified for any analysis in this paper.