TEFS: REJECTED-RECORD NONINTERFERENCE IN TYPED HYBRID DOCUMENT QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Hybrid document-QA runtimes select among records returned by readers, structural parsers, visual models, and symbolic operators. We identify a missing integrity boundary: a completed record that will later be rejected can still influence candidate-dependent computation and change which admissible record is ultimately selected. We formalize *rejected-record noninterference* (RNI) and characterize it through invariance of the selector-visible view to rejected records. TEFS enforces completed-record admissibility before this view is formed, preventing rejected candidates from affecting downstream features, rankings, or selector state. Natural replay and paired live-return experiments show that rejected records can alter admissible selections even when they are never selected themselves, while controlled return faults distinguish this boundary from call-time authorization and partial post-execution checks. Under its trusted-request and trusted-wrapper assumptions, TEFS guarantees contract-consistent participation rather than semantic correctness, contract inference, or deployment-wide fault prevalence.

1 INTRODUCTION

A candidate that will ultimately be rejected can still change which admissible answer wins. Typed hybrid document-QA runtimes combine heterogeneous producers, while safeguards usually act either before execution (authorizing calls) or after selection (filtering outputs). Neither protects the intermediate selector-visible computation. We study this post-execution, pre-selection integrity boundary.

Consider the typed request DIFFERENCE$(+0.180, +7.020)$, whose authorized arithmetic implementation computes $7.020 - 0.180 = 6.840$. A general reader returns 6.839 with evidence, while an authorized calculator returns 6.840. A table specialist instead copies the visible row label 6 and cites that region. Its trusted-wrapper trace records copying rather than the requested difference. The record is grounded and executable, but it does not implement the declared operation. Admissibility therefore requires that the recorded interface and canonical arguments—together with request, document, coordinate, version, registration, and provenance bindings—conform to the typed request and fixed policy (Figure 1A).

Pre-execution authorization is also insufficient (Figure 1B). Retrieval controls evidence, grounding links an answer to a source, routing estimates utility, and tool authorization checks an intended call. However, an authorized invocation may still be followed by a cache, adapter, or producer return belonging to another operation, argument tuple, document region, version, or registration. The systems compared here do not jointly validate these properties on the completed returned record before selection. Authorization establishes that a call was permitted; it does not establish that the returned record belongs to that call.

Final-output masking is too late. In the separate schematic selector witness of Figure 1C, admissible reader c_1 and calculator c_2 have base scores .78 and .74. Rejected copied-value record c_r returns 6.840, agrees with c_2's 6.840, and adds .10 to c_2's agreement feature. Late computation therefore raises c_2 to .84 and selects it; masking c_r afterward cannot restore the uninfluenced winner c_1. TEFS hides c_r before feature computation, leaving .78 versus .74. This example concerns causal influence over selection, not whether the uninfluenced winner is factually correct. *Candidate-dependent computation* comprises any feature, score, state transition, or ranking that depends on the candidate

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



Figure 1: Three failures separated by candidate admissibility. Grounding does not establish that returned execution metadata conforms to the requested operation (A); authorization does not establish the returned binding (B); and post-feature masking lets a rejected non-winner alter the admissible winner (C). Panel C is a separate selector witness, not a second arithmetic trace of the document in A. TEFS admits before selector-visible computation.

sequence. Exclusion must precede all such computation, not merely prevent an inadmissible final choice.

TEFS (Typed Eligibility-First Selection) makes completed-record admissibility a precondition for selector participation. Trusted wrappers record invocations; the monitor admits only kernel-owned, mutation-protected Candidate Records before candidate-dependent computation. This separates *availability* (can a producer return?), *admissibility* (may the record participate?), and *utility* (which admitted record should win?). TEFS constrains admissibility, not utility. Under its trusted-request and trusted-wrapper assumptions it establishes invocation admissibility and binding integrity, not semantic result validity: an admitted answer may still be false.

Our contributions are:

- We identify completed-record selector participation as a distinct integrity boundary, formalize *rejected-record noninterference*, and characterize its arbitrary-selector guarantee by selector-view invariance.
- We instantiate the property with a pre-selection admission kernel that binds operation, canonical arguments, evidence, document/version state, registration, and provenance before candidate-dependent computation.
- We show 82 admitted-winner changes despite zero rejected winners in a 6,290-request non-injected replay and five further changes in each of three repeated 629-request live-return runs; controlled faults and transfer isolate the boundary from authorization and utility.

2 Problem Boundary and Related Work

Reference monitors, noninterference, and postconditions. Reference monitors and complete mediation require protected operations to pass an analyzable enforcement point (Anderson, 1972; Saltzer & Schroeder, 1975); classical noninterference asks whether one class of actions can change what another observes (Goguen & Meseguer, 1982). TEFS claims neither principle as new. ToolGate, the closest post-execution analogue, verifies Hoare-style postconditions before an individual tool result commits a typed symbolic-state update (Liu et al., 2026). TEFS instead protects a shared selector-visible view: a rejected multi-producer record may change another admissible record without itself being committed or emitted. The contribution is this selection-specific protected object,

RNI objective, and enforcement placement; a postcondition checker can be one implementation of the returned-record gate.

Table 6 in Appendix A summarizes the closest mechanisms by protected object; the comparison is not a capability ranking.

Database provenance traces explain how results derive from execution (Cheney et al., 2008); recent agent-provenance work broadens this view to retrieval, tools, memory, and actions (Wang et al., 2026). TEFS uses provenance as one conjunct of admission rather than treating provenance alone as proof of authorization or correctness. Generic proposer–verifier–commit–reject runtimes are complementary (Srinivasan, 2026).

Evidence acquisition, routing, and transactions. Visual document retrieval exposes pages, regions, or elements (Yu et al., 2024; Sun et al., 2025; Li et al., 2025; Chen et al., 2026a), while hybrid systems add structural or symbolic specialists (Cao et al., 2025; Abhyankar et al., 2025; Zhang et al., 2025; Yang et al., 2026). These mechanisms improve evidence access or utility, but grounding does not establish that execution metadata matches the requested operation. Learning-to-defer and expert routing optimize expected utility (Mozannar & Sontag, 2020; Verma et al., 2023; Mao et al., 2025); they answer which admissible option is useful, not whether a completed return belongs to the request. Cordon validates task-level result lineage and stages external effects before commit (Chen et al., 2026c); TEFS addresses the complementary read-only boundary after each producer return and before selector participation. SkVM compiles skills across heterogeneous model–harness targets (Chen et al., 2026b); TEFS constrains the returned-record domain below such integration machinery. Programmed adapters in our experiments localize boundaries; they do not establish capability limits or rank external systems.

3 Candidate Admissibility and the TEFS Kernel

Threat model and trusted base. The application supplies an assumed-correct, schema-valid typed request $\kappa = (q, o, \alpha, d, v_\kappa)$ containing request identity q, logical operation o, canonical arguments α, document binding d, and request-side versions v_κ, together with fixed policy P. The evaluated operations are read-only. Producers, measured profiles, compiler transformations, caches, and selectors are not trusted to grant authority. The trusted computing base is the typed-request schema validator, policy loader, operation canonicalizers, registered wrappers, kernel-controlled record store, range check, and atomic emitter. A cooperative wrapper observes each actual invocation and writes execution metadata that the producer cannot modify. The guarantee targets integration faults, stale state, and misbinding despite producer-controlled answer and evidence payloads; it does not validate those payloads semantically or cover a forged wrapper, incorrect typed request, or pre-kernel effects.

Candidate Record. For answer a, evidence E, and runtime metadata, a Candidate Record c contains the concrete schema summarized in Table 1. The record distinguishes logical operation $o(c)$ from implementation class $m(c)$ and records the canonical arguments $\beta(c)$ actually supplied by the wrapper. Document, coordinate, and version fields travel with the result rather than remaining implicit in controller state. "Mutation-protected" means a kernel-controlled object within the evaluated runtime; authenticated cross-service serialization is outside the present evidence.

Common admission interface. Every producer class, including a general reader, passes one admission interface that applies a common record contract with operation-specific policy and canonicalization:

$$\begin{aligned}\mathrm{Adm}_P(\kappa, c) = \ & \mathrm{Schema}(c) \wedge \mathrm{Success}(c) \\ & \wedge\, o(c) = o \wedge m(c) \in \mathcal{I}_P(\kappa) \wedge \beta(c) \equiv_o \alpha \\ & \wedge\, \mathrm{ReqBound}(c, \kappa, P) \wedge \mathrm{DocBound}(c, d) \\ & \wedge\, \mathrm{VersionBound}(c, v_\kappa) \wedge \mathrm{ProvIntact}(c).\end{aligned} \tag{1}$$

Here $\mathcal{I}_P(\kappa)$ is the set of implementations authorized by the fixed policy for request κ, $\equiv_o$ is the operation-specific canonical-argument relation, and the binding predicates compare complete identities rather than request ID alone. Table 1 maps each predicate to its required fields; expected values

Table 1: Candidate Record fields and the binding predicates in Equation 1. Missing required fields fail admission; operation and canonical arguments are checked separately. The emitter projects from one selected record.

Group	Required fields
Result	answer; evidence payload and type
Operation	logical operation; implementation; canonical executed arguments
REQBOUND	request ID and plan identity/signature match the trusted request and plan; execution status is successful
DOCBOUND	document ID and digest/version match the plan; requested page agrees with present bindings; evidence type, required region fields, and coordinate frame meet the operation schema
VERSIONBOUND	policy, operation catalog/schema, producer profile, preprocessing, coordinate transform, wrapper schema, and plan versions match current registration
PROVINTACT	wrapper invocation, producer/registration and unique record identities, runtime ownership, and kernel fingerprint remain intact

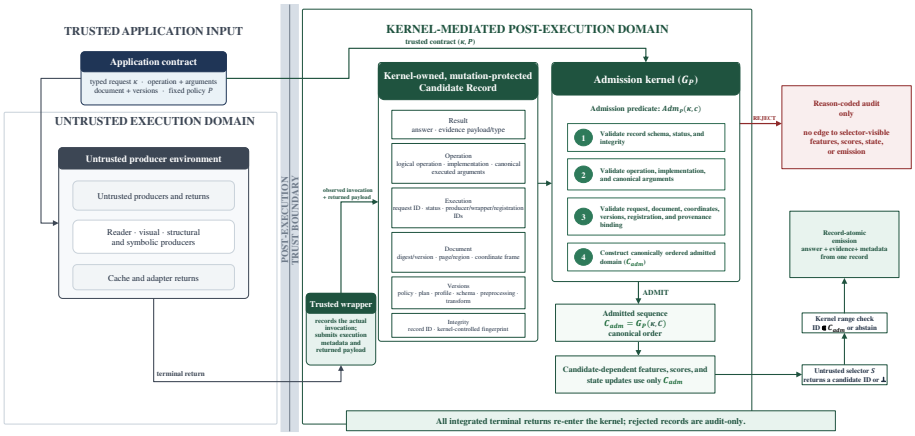


Figure 2: The TEFS placement contract. Every completed producer return re-enters a trusted admission boundary before candidate-dependent computation. Rejected records are audit-only; admitted records alone form the selector-visible view. The kernel range-checks the selector output and atomically emits one record. Table 1 gives the fields checked at admission.

come from the trusted request, policy, plan, and registration rather than the returned payload. A policy may authorize $m = \texttt{general_semantic_answer}$ as an implementation of o; its wrapper must invoke that interface with the current (o, α, d, v_κ) and record the same fields as any specialist. This removes the previous reader exception. For an opaque model, the predicate establishes invocation of the authorized interface, not its latent reasoning process or factual correctness. Throughout, *rejected* means any record with $\neg\,\mathrm{Adm}_P(\kappa, c)$. When the reason matters, we name the subclass explicitly—failed/malformed, operation or argument incompatible, binding/version mismatched, or provenance/integrity invalid—rather than using these terms interchangeably.

For an input sequence C, let $\prec$ be a fixed total order on kernel-issued record identities, independent of arrival order; the trusted store rejects identity collisions. A collision is an append failure: it cannot overwrite, mutate, or invalidate the previously stored record. Admission returns

$$G_P(\kappa, C) = \mathrm{sort}_\prec\left(\langle c \in C : \mathrm{Adm}_P(\kappa, c)\rangle\right). \tag{2}$$

Let $C_{\mathrm{adm}} = G_P(\kappa, C)$, selector state be σ, coupled randomness be ω, and $z(C_{\mathrm{adm}})$ contain only features of C_{adm}. A range-safe selector has interface $S(C_{\mathrm{adm}}, z(C_{\mathrm{adm}}), \omega, \sigma) \to (y, \sigma')$, where $y \in C_{\mathrm{adm}} \cup \{\perp\}$. Neither feature nor selector code can observe raw C, rejected-record metadata, or arrival-order side channels. If $y \in C_{\mathrm{adm}}$, the emitter atomically projects answer, evidence, and metadata from y; $y = \perp$ or $C_{\mathrm{adm}} = \emptyset$ invokes the application's declared abstention policy. Admission uses no gold answer or utility score.

Definition 1 (rejected-record noninterference). Let Obs contain the selected record identity or declared abstention, emitted answer–evidence tuple, and post-selection state; it explicitly excludes audit-only logs, counters, and timing. Write $\mathcal{O}(C) = \mathrm{Obs}(\mathrm{TEFS}_P[S](\kappa, C; \sigma, \omega))$. A runtime satisfies rejected-record noninterference when, for every completed sequence C and additional se-

Table 2: Complete natural 2×2 placement replay on 6,290 requests, with identical archived returns and a frozen agreement selector. Rejected visible counts successful rejected returns entering features, not violations at emission. The 48 visible under authorization alone change no legal score or winner here. All arms apply the same late range check; controlled binding faults appear only in Appendix A.

Call auth. / pre-feature admission	Candidate calls	Rejected visible	Score-changed requests	Top / legal-choice changes	EM
Off / off	31,450	1,411	1,042	274 / 82	50.40
On / off	28,934	48	0	0 / 0	50.30
Off / on	31,450	0	0	0 / 0	50.30
On / on	28,934	0	0	0 / 0	50.30

quence R,

$$[\forall r \in R, \neg \operatorname{Adm}_P(\kappa, r)] \Longrightarrow \mathcal{O}(C \parallel R) = \mathcal{O}(C), \tag{3}$$

provided generation has finished, both executions begin with identical state and coupled randomness, and rejected records have not already changed shared state or external effects.

Theorem 1 (selector-view characterization). Let $V_M(\kappa, C; \sigma, \omega)$ contain the ordered eligible identities, candidate fields, features, and pre-selection state visible to a runtime's selector, but not audit-only output. Assume range-safe selection, record-atomic emission, and identical initial state and coupled randomness. For *every* selector allowed to update observable post-selection state, rejected-record noninterference holds if and only if V_M is invariant to appending rejected records. *Proof.* Equal views give the same selector identical inputs, hence identical selected identity, emitted projection, and state. If two views differ, a range-safe selector can distinguish them and write one bit into observable post-selection state, even when both views force abstention. With an admitted identity it can instead choose that identity versus abstention; with two, it can change the legal winner. A fixed selector may ignore a changed view, so necessity is for the arbitrary-selector guarantee.

Corollary (pre-feature admission). Stable canonical ordering and mutation-protected records give $G_P(\kappa, C \parallel R) = G_P(\kappa, C)$ for rejected R. Computing every selector-visible feature and state only from this admitted sequence constructs the invariant view of Theorem 1. Final masking cannot restore that guarantee after a rejected record has changed the view.

Kernel path. The kernel canonicalizes the request, verifies Equation 1, exposes only $G_P(\kappa, C)$, checks the selector's returned identity against that sequence, and atomically projects answer and evidence from the selected mutation-protected record. Cache hits take the same path: the kernel revalidates the returned record's operation, arguments, document, coordinates, versions, and current registration. A mismatch discards the cached value and invokes the currently registered implementation.

Producer integration. Reader, visual, structural, and symbolic paths may execute independently, but each integrated terminal path must return through the same Candidate Record and admission interface. Profiles and compilation may remove or rank policy-authorized implementations; they cannot add authority.

4 Experimental Methodology

We use MMDocBench (Zhu et al., 2024), DOGR (Zhou et al., 2025), and TABVERSE (Ahsan et al., 2026). Typed mappings, policies, candidate contents, and selector settings are frozen before outcome scoring. Admissibility labels are deterministic conditional on the frozen mapping, policy, canonicalizers, and runtime fields; they are not answer-correctness annotations. The policy declares application-supplied direct operations and is not tuned for answer accuracy. We keep evidence classes separate: unmodified frozen selectors measure observed non-injected incidence; the 6,290-request five-source replay isolates placement; and rotated operations, cache donors, stale bindings, and factorial endpoints are controlled fault probes rather than incidence evidence. The ten-operation TABVERSE protocol was frozen before target inspection; the MMDocBench/DOGR stress pools and the fully live pipeline are post-hoc studies whose gold-free decisions were frozen before scoring.

Statistical protocol. Unless noted otherwise, accuracy intervals are paired percentile-bootstrap intervals with 10,000 resamples over natural source-image, document-image, or released-table groups. Archived experiment artifacts retain grouping, seed, and hash manifests. Live-pipeline intervals condition on one frozen execution. Three full greedy primary executions and three additional nonzero-temperature executions give separate descriptive run-level means and standard deviations, not precise run-level intervals. The archived greedy selector used temperature 0, top-$p = 1$, a fixed seed, and at most 16 completion tokens; the sensitivity runs use temperature 0.2 and three distinct seeds for generation and selection. Appendix A gives the matched-repeat audit. Local-controller overhead uses 5,000 paired resamples.

RQ1: Can rejected candidates affect legal selections without being selected? Table 2 summarizes four placements that replay the same five-record sequence and agreement selector on 6,290 TABVERSE requests. Outcomes are rejected records visible to features, requests whose admitted scores change, and selected admitted-record changes.

RQ2: How often do frozen selectors choose incompatible returns? Four unmodified frozen selectors measure selector-specific incidence. For a pre-admission replay, an already frozen pairwise decision is reused only when its complete two-record input remains admitted; a singleton admitted sequence returns its sole record without a pairwise call, and an empty sequence abstains. Candidate generation and model revision are paired, but selector call counts are not. EM and strict joint answer–evidence accuracy measure utility after this projection.

RQ3: Which returned-record checks are required? Appendix A separates returned operation/arguments from complete binding. Rotated-operation returns preserve answer, success, and method family while changing the recorded logical operation. Separately, 3,034 eligible cache substitutions use a source-valid donor from another operation on the same table, so exact operation-and-argument revalidation rejects them by construction. The constructor starts from 3,774 possible pairs and excludes 740 failed or malformed source donors. Nine one-at-a-time mutations hold request ID fixed while changing arguments, document, coordinates, policy, producer, or a relevant version. Programmed adapters localize their configured enforcement points; they are diagnostic configurations, not competitive baselines or capability limits. A separate 2^6 design tests these six factors on disjoint endpoint strata.

RQ4: Does the returned-record boundary transfer? We test complete-candidate admission on DOGR with six frozen readers, on the chronologically untouched ten-operation TABVERSE transfer, and in a fully live four-producer Candidate Record pipeline. These protocols distinguish cross-dataset enforcement from utility of the downstream selector.

RQ5: What do interface conformance and cost establish? TEFS assumes a typed runtime interface; automatic natural-language contract induction is upstream and outside its core claim. Ten typed families test kernel conformance, not task accuracy. Cost uses 500 randomized paired local kernel executions and a separate 500-request Qwen3.6 profile.

5 RESULTS

5.1 RQ1: REJECTED NON-WINNERS CAN CHANGE LEGAL SELECTIONS

Post-feature or absent enforcement lets 1,411/31,450 records enter selector features. Those participations change 2,705 admitted-candidate score entries on 1,042/6,290 requests, alter the legal top set on 274 requests, and change the selected admitted record on 82, although no inadmissible record wins. Raw/late enforcement scores 50.397 EM versus 50.302 for TEFS ($+0.095$ [-0.159,0.350]); the system property is influence, not a guaranteed accuracy direction. A final-winner mask cannot undo this influence. Pre-feature TEFS makes rejected records observationally absent under Theorem 1, and all four influence counts become zero. An evaluation counting only inadmissible winners would miss all 82 legal-winner changes. The theorem and three-candidate witness establish why placement, rather than only the admission predicate, is necessary. Pre-execution authorization leaves 48 successful but malformed direct-operation returns visible to features; they do not change

an admitted score or winner for this selector, which has the same 50.302 EM as TEFS. Visibility is not observed influence.

Two archived traces show both accuracy directions. On `009a1407:count_rows`, admitted reader answers are 15/17/16/17; the rejected malformed specialist says 16. Late masking selects an admitted 16, while early admission selects 17 (gold 17). Conversely, on `0c7fa7fc:count_columns`, admitted answers are 8/7/7/8 and the rejected specialist says 8. Late masking selects admitted 8 (gold 8), but early admission selects 7. Neither rejected record wins. Across all 82 changed winners, early admission repairs 30 answers, harms 36, and leaves both wrong in 16.

Theorem 1 predicts a selector-class split. For a rejected *non-winner*, candidate-separable scoring $s_i = f(c_i)$ has no cross-candidate channel: adding c_r cannot change another candidate's score, although c_r itself may still be selected and late-masked. Set-dependent selection $S(C)$ or scoring $s_i = f(c_i, C)$ can expose such a channel through agreement, consensus, attention, or joint prompting. Table 3 tests both classes on the same return pool.

Table 3: Mechanism test on an internal table split (126 test tables, 1,260 requests). The null rows are predicted controls for candidate-separable scoring, not failed replications of the set-dependent mechanism.

Selector	Candidate interaction	Legal winners changed	EM early / late
Agreement	Set-dependent	16 / 1,260	48.02 / 48.33
Reader priority	Separable	0 / 1,260	43.25 / 43.25
Development-fit utility	Separable	0 / 1,260	64.68 / 64.68
Calibrated consensus	Effective separable (w=0)	0 / 1,260	64.68 / 64.68

Development tuning selected consensus weight zero from $\{0, 0.25, 0.5, 1, 2\}$, so the calibrated-consensus null collapses to the separable case. The positive side is not limited to hand-coded agreement: RQ4 jointly presents candidate sets to a prompted Qwen selector and finds five changed admissible winners in each of three greedy live-return replays, with 2–10 changes across three temperature-0.2 runs. This establishes the cross-candidate mechanism for a learned joint selector without implying that every set-dependent selector must change on every pool. The internal split is used specifically to diagnose the selector-class mechanism.

An independently coded generic early postcondition gate on raw record fields reproduces TEFS's admitted view and winner on all 6,290 requests (Appendix A). Thus TEFS is not claimed as the unique algorithmic realization of RNI; the contribution is the protected selector view, the RNI property, and the placement contract that realizes it.

5.2 RQ2: Observed Incidence and the Compliance–Utility Trade-off

Table 4: Observed non-injected, selector-specific incidence. Frozen pairwise decisions are reused only for intact admitted pairs; singletons follow the declared rule. Pairwise deltas are $+1.89$ EM [1.43,2.35] and $+1.41$ joint [0.93,1.91] (n = 4,338 and 3,466). Utility rows use the secondary no-percent-rescaling scorer; gated-minus-raw EM intervals are -1.16 [-2.31,-0.17] and -0.99 [-1.98,-0.17].

Frozen selector	Requests	Specialist choices	Inadmissible choices	EM raw→pre-adm.	Joint raw→pre-adm.
Pairwise Qwen2.5-VL-7B	4,338	514	363	68.67→70.56	61.71→63.13
Cross-fitted learned	4,338	245	0	71.76→71.76	64.60→64.60
Matched one-order utility	605	122	18	73.39→72.23	–
Matched two-order utility	605	59	14	70.74→69.75	–

Blindly prioritizing 2,708 available structural proposals on MMDocBench collapses EM/joint from 69.59/62.03 to 33.54/33.35; the supplied-request contract admits 263 and reaches 71.76/64.60. This is a failure-mode control, not a competitive selector.

The pairwise VLM chooses 363 inadmissible specialists: 8.37% of 4,338 requests and 70.62% of its 514 specialist choices. The learned selector is an informative null with zero inadmissible choices. The utility selectors choose 18 and 14 inadmissible records; 14/18 and 12/14 of those answers are EM-correct, and 9 and 7 uniquely repair an incorrect reader. Enforcing the contract

lowers their secondary-scorer EM by 1.16 points $[-2.31,-0.17]$ and 0.99 points $[-1.98,-0.17]$, respectively. TEFS is justified when operation and provenance compliance are requirements; it does not maximize unconstrained EM. Yet both wrapped selectors remain useful relative to their same-run 67.44 EM reader: one-order TEFS reaches 72.23 ($+4.79$ [2.64,7.11]) and two-order TEFS reaches 69.75 ($+2.31$ [0.66,3.97]), each with zero incompatible choices. Thus admission constrains, rather than replaces, utility selection.

Pairwise pre-admission replay reuses frozen decisions only for unchanged admitted pairs; otherwise it applies the singleton rule. It retains 151 specialist choices and removes all 363 violations, but does not hold selector call count fixed. Appendix A reports the reason-code audit.

Auditable trade-off cases. On `CharXiv_13_3`, an inadmissible specialist returns the reference count 5 while the reader returns 6; its wrapper record reports an unknown executed operation, so strict admission preserves the wrong reader answer. On TABVERSE request `009a1407:count_columns`, the successful but count-incompatible specialist returns 8 while the reader and reference are 7; admission preserves 7. These cases make the distinction operational: availability does not imply admissibility, and admissibility does not imply utility. All three must be reported separately.

5.3 RQ3: Returned-Record Checks Isolate Distinct Faults

The natural replay distinguishes call-time authorization from TEFS in feature visibility, but not selected winner or EM. Binding-fault effects are tested separately by controlled probes (Appendix 8): method-only checking admits 3,082 rotated-operation returns; call-time authorization exposes 3,034 cross-operation cache donors and selects ten; operation-and-argument revalidation rejects all 3,034. Holding request ID fixed, nine stale-binding mutations pass an ID-only check but not full revalidation. None estimates natural fault prevalence.

Factorial controls. All 64 programmed configurations confirm that each factor blocks its designed failure. Without exact arguments, 263 misbindings cause 92 answer harms; without mutation protection or atomic emission, 10/10 endpoint attacks pass. Atomic emission raises joint accuracy from 64.08 to 64.60 at fixed 71.76 EM.

5.4 RQ4: Cross-Dataset Transfer and Live Execution

Across three non-injected boundary-exposing pools, 4,292/7,747 completed specialist records are incompatible with their supplied requests: 2,445/2,708 on MMDocBench, 484/594 on DOGR, and 1,363/4,445 on TABVERSE. The MMDocBench and DOGR pools are designed post-hoc stress tests; only the TABVERSE protocol was frozen before target inspection. These ratios therefore show that availability and admissibility differ across datasets and producer types, not production prevalence.

On DOGR, every crop is grounded in a supplied ROI, but only 110/594 completed crops implement a direct-transcription request; the other 484 are visible text for reasoning, mathematics, or estimation questions. Supplied typed admission beats unconditional crop priority by 21.17–40.50 EM points across six frozen readers, with all source-image-grouped intervals excluding zero. Under the proper eligible-only protocol it improves five native readers by 2.83–4.00 points and is neutral for Qwen3.6. Thus grounding identifies a source region but does not establish that copying that region implements the requested operation.

In Table 5(a), typed admission removes 1,363 violations and gains $+3.15$ [2.59,3.69] EM points over availability priority, but remains 8.43 below reader-only: enforcement does not replace utility selection. Method-only checking admits all 3,082 rotated returns; full matching admits none. The live pipeline makes 1,764 producer and 597 selector calls and reaches 55.01 EM versus 44.83 for same-run Qwen2.5-VL-7B records ($+10.17$ [6.20,14.15]); this is a pipeline-feasibility result, not an isolated admission effect.

Paired admission replay on live returns. Across three repeats of the *same* 629 requests with fixed greedy seed and model revisions, paired on/off replay over identical successful returns changes five admissible winners per run and yields 17, 17, and 18 incompatible off-arm choices. On/off

Table 5: Transfer and live-system evidence. Panel (a) uses candidate ledgers frozen before target opening; availability is format-valid, not authorized. Panel (b) executes two VLMs, OCR, grid, admission, and a range-checked prompted Qwen2.5-VL-3B selector. *Producer EM is attempt-domain; final EM spans all requests and is a pipeline-level result, not an isolated admission effect.

(a) Chronologically untouched TABVERSE transfer: 629 tables, ten operations, 6,290 requests.

Policy	EM	Specialist selected	Incompatible selected	Repairs / harms	Δ vs. reader [95% CI]
Reader only	44.56	0	0	0 / 0	+0.00 [0.00,0.00]
Availability priority	32.99	4,445	1,363	285 / 1,013	−11.57 [−12.50,−10.64]
Typed admission	36.14	3,082	0	193 / 723	−8.43 [−9.17,−7.71]
Rotated return, method only	27.34	3,082	3,082	1 / 1,084	−17.22 [−18.00,−16.45]
Rotated return, full typed	44.56	0	0	0 / 0	+0.00 [0.00,0.00]
Eligible oracle	47.63	193	0	193 / 0	+3.07 [2.67,3.48]

(b) Fully live four-producer Candidate Record pipeline: 629 requests.

Profile / output	Producer attempts	Successful records	Selected	Selector calls	EM*	Latency P50 / P95 (s)
Qwen2.5-VL-7B reader	629	611	45	–	44.83	0.738 / 1.513
Qwen2.5-VL-3B reader	569	515	121	–	35.68	0.482 / 1.112
PP-OCRv5 targeted cells	377	323	279	–	37.67	0.314 / 0.700
Deterministic pixel grid	189	187	184	–	93.12	0.070 / 0.228
TEFS + live selector	**1,764**	**1,636**	**629**	**597**	**55.01**	**0.928 / 2.027**

EM is 55.01/54.69, 55.17/54.85, and 55.01/54.69 (on-arm mean 55.06%, sample SD 0.09 points). Three temperature-0.2 seeds retain 2–10 changed winners and 8–18 incompatible choices with only a 0.05-point mean EM gap. These repeated executions quantify run-to-run robustness on a fixed request collection; Appendix A gives abstentions, uncertainty, and audits. No successful, well-formed post-return binding mismatch occurs in these live runs; controlled binding faults therefore establish prevention, not deployment prevalence.

5.5 RQ5: Interface Conformance and Cost

Ten typed operation families pass 10/10 deterministic checks. A full-ledger scan flags 648 requests, none containing the 82 winner changes (Appendix A). Across 500 paired local executions, median controller latency rises from 0.124 to 1.400 ms: +1.277 ms [1.257,1.297], excluding scaling, serialization, contention, and throughput.

6 Limitations

TEFS assumes correct typed requests, fixed policy, faithful canonicalizers, and trusted wrappers; malicious wrappers can forge fields. Contract induction is upstream and can fail (208/489 false admissions in a post-hoc DOGR diagnostic), and the remaining contracts lack full human audit. TEFS does not guarantee semantic execution, factual correctness, or evidence entailment. The live runs contain no successful, well-formed post-return binding mismatch, so controlled probes establish prevention rather than deployment prevalence. Strict policy can reject correct answers or empty the pool; RNI excludes generation-time shared state and side channels; local mutation protection is not cross-service authentication; and cost measurements omit throughput and contention.

7 Conclusion

Hybrid selection introduces an integrity boundary after producer return but before candidate-dependent computation. Rejected non-winners changed 82 admissible winners in the archived replay and again altered legal winners under a prompted neural selector in repeated live-return runs. RNI characterizes the required selector-view invariance, and TEFS realizes one implementation by enforcing completed-record admissibility before candidate interaction. More broadly, selection integrity depends on which records are allowed to influence the selector, not only which records may ultimately be emitted. The guarantee remains contract-consistent participation, not correct typing, semantic correctness, or utility.

AI USE STATEMENT

Generative AI assisted with experimental implementation, result interpretation, controlled synthetic fault construction, the selector-view theorem and proof, related-work discovery, drafting and polishing text. The authors take responsibility for independently checking AI-assisted claims, proofs, citations, code, figures, and numerical summaries against the cited sources and archived experimental artifacts, and for the final manuscript.

REPRODUCIBILITY STATEMENT

The main text specifies the trust model, admission predicate, datasets, selector protocols, uncertainty estimates, and principal denominators. Appendix A reports all seven placement arms, controlled-fault denominators, the paired live-selector replay, repeat and cache audits, and component costs. Controlled faults are explicitly separated from observed non-injected incidence throughout the paper.

REFERENCES

Nikhil Abhyankar, Vivek Gupta, Dan Roth, and Chandan K. Reddy. H-STAR: LLM-driven hybrid SQL-text adaptive reasoning on tables. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8841–8863, 2025. doi: 10.18653/v1/2025.naacl-long.445. URL https://aclanthology.org/2025.naacl-long.445/.

Momina Ahsan, Sarfraz Ahmad, Ming Shan Hee, Roy Ka-Wei Lee, and Preslav Nakov. TABVERSE: Benchmarking cross-format table understanding in LLMs and VLMs, 2026. URL https://arxiv.org/abs/2606.09578.

James P. Anderson. Computer security technology planning study. Technical Report ESD-TR-73-51, Electronic Systems Division, Air Force Systems Command, 1972. URL https://apps.dtic.mil/sti/citations/AD0758206.

Ruisheng Cao, Hanchong Zhang, Tiancheng Huang, Zhangyi Kang, Yuxin Zhang, Liangtai Sun, Hanqi Li, Yuxun Miao, Shuai Fan, Lu Chen, and Kai Yu. NeuSym-RAG: Hybrid neural symbolic retrieval with multiview structuring for PDF question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6211–6239, 2025. doi: 10.18653/v1/2025.acl-long.311. URL https://aclanthology.org/2025.acl-long.311/.

Guanhua Chen, Chuyue Huang, Yutong Yao, Shudong Liu, Xueqing Song, Lidia S. Chao, and Derek F. Wong. From scenes to elements: Multi-granularity evidence retrieval for verifiable multimodal RAG. *arXiv preprint arXiv:2605.15019*, 2026a. doi: 10.48550/arXiv.2605.15019. URL https://arxiv.org/abs/2605.15019.

Le Chen, Erhu Feng, Yubin Xia, and Haibo Chen. SkVM: Revisiting language VM for skills across heterogenous LLMs and harnesses, 2026b. URL https://arxiv.org/abs/2604.03088.

Zheng Chen, Hanqing Liu, Duling Xu, Dong Dong, Jialin Li, Bangzheng Pu, and Jidong Zhai. Cordon: Semantic transactions for tool-using LLM agents, 2026c. URL https://arxiv.org/abs/2606.17573.

James Cheney, Umut A. Acar, and Amal Ahmed. Provenance traces, 2008. URL https://arxiv.org/abs/0812.0564.

Joseph A. Goguen and José Meseguer. Security policies and security models. In *1982 IEEE Symposium on Security and Privacy*, pp. 11–20. IEEE Computer Society, 1982. doi: 10.1109/SP.1982.10014.

Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms, 2020. URL https://arxiv.org/abs/2006.14171.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. RegionRAG: Region-level retrieval-augmented generation for visual document understanding. *arXiv preprint arXiv:2510.27261*, 2025. doi: 10.48550/arXiv.2510.27261. URL https://arxiv.org/abs/2510.27261.

Yanming Liu, Xinyue Peng, Jiannan Cao, Xinyi Wang, Songhang Deng, Jintao Chen, Jianwei Yin, and Xuhong Zhang. ToolGate: Contract-grounded and verified tool execution for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9653–9684, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.470. URL https://aclanthology.org/2026.findings-acl.470/.

Dongsheng Ma, Jiayu Li, Zhengren Wang, Yijie Wang, Jiahao Kong, Weijun Zeng, Jutao Xiao, Jie Yang, Wentao Zhang, Bin Wang, and Conghui He. CiteVQA: Benchmarking evidence attribution for trustworthy document intelligence. *arXiv preprint arXiv:2605.12882*, 2026. doi: 10.48550/arXiv.2605.12882. URL https://arxiv.org/abs/2605.12882.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable H-consistency and strong guarantees for learning to defer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43035–43066. PMLR, 2025. URL https://proceedings.mlr.press/v267/mao25c.html.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Jerome H. Saltzer and Michael D. Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975. doi: 10.1109/PROC.1975.9939. URL https://doi.org/10.1109/PROC.1975.9939.

Vasundra Srinivasan. A methodology for selecting and composing runtime architecture patterns for production LLM agents, 2026. URL https://arxiv.org/abs/2605.20173.

Yubo Sun, Chunyi Peng, Yukun Yan, Shi Yu, Zhenghao Liu, Sen Mei, Chi Chen, and Maosong Sun. VisRAG2.0: Mitigating visual hallucinations via evidence-guided multi-image reasoning in visual retrieval-augmented generation. *arXiv preprint arXiv:2510.09733*, 2025. doi: 10.48550/arXiv.2510.09733. URL https://arxiv.org/abs/2510.09733.

Venkata Kesav Venna, Sai Madhusudan Gunda, Jyothi Swaroopa Jinka, Hrithik Sagar Rachakonda, Anirudh Srinivasan, and Ravi Kiran Sarvadevabhatla. M3Grounder: Mask-based multi-span and multi-granular grounding for document qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23685–23695, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Venna_M3Grounder_Mask-Based_Multi-Span_and_Multi-Granular_Grounding_for_Document_QA_CVPR_2026_paper.html.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11415–11434. PMLR, 2023. URL https://proceedings.mlr.press/v206/verma23a.html.

Yiqi Wang, Jiaqi Zhang, Taotao Cai, Zirui Liu, Qingqiang Sun, Zequn Sun, Zhangkai Wu, Manqing Dong, Mingkai Zheng, Xuefei Yin, and Yanming Zhu. From agent traces to trust: A survey of evidence tracing and execution provenance in LLM agents, 2026. URL https://arxiv.org/abs/2606.04990.

Cailin Winston, Claris Winston, and René Just. Solver-aided verification of policy compliance in tool-augmented LLM agents, 2026. URL https://arxiv.org/abs/2603.20449.

Jiashu Yang, Chi Zhang, Abudukelimu Wuerkaixi, Xuxin Cheng, Cao Liu, Ke Zeng, Xu Jia, and Xunliang Cai. AutothinkRAG: Complexity-aware control of retrieval-augmented reasoning for image-text interaction, 2026. URL https://arxiv.org/abs/2603.05551.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*, 2024. doi: 10.48550/arXiv.2410.10594. URL https://arxiv.org/abs/2410.10594.

Junwen Zhang, Pu Chen, and Yin Zhang. TableMoE: Neuro-symbolic routing for structured expert reasoning in multimodal table understanding, 2025. URL https://arxiv.org/abs/2506.21393.

Yinan Zhou, Yuxin Chen, Haokun Lin, Yichen Wu, Shuyu Yang, Zhongang Qi, Chen Ma, and Li Zhu. DOGR: Towards versatile visual document grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3596–3606, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_DOGR_Towards_Versatile_Visual_Document_Grounding_and_Referring_ICCV_2025_paper.html.

Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. MMDocBench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*, 2024. doi: 10.48550/arXiv.2410.21311. URL https://arxiv.org/abs/2410.21311.

Genliang Zhu and Chu Wang. Intent-governed tool authorization for AI agents, 2026. URL https://arxiv.org/abs/2606.22916.

A Additional Protocol and Audit Details

Table 6: Boundary positioning of the closest mechanism classes. "Shared view" means that candidate-dependent selector computation itself is inside the protected object. This is a protected-object comparison, not a capability ranking; the mechanisms can be composed.

Mechanism	Primary enforcement point	Shared view?	Protected object / objective
Grounding / citation (Ma et al., 2026; Venna et al., 2026)	Evidence/source attribution	No	Answer–evidence provenance
Call authorization / action masks (Huang & Ontañón, 2020; Zhu & Wang, 2026; Winston et al., 2026)	Intended call or final action	No	Permitted invocation / choice
ToolGate (Liu et al., 2026)	Tool result before symbolic-state commit	Not its target	Typed tool-result state transition
TEFS / RNI	Completed record before candidate interaction	**Yes**	Selector-view noninterference

Evaluation and uncertainty. The natural five-source placement replay uses 6,290 TABVERSE requests and 31,450 completed records. Accuracy intervals use 10,000 paired percentile bootstrap resamples over the natural grouping unit (source image, document image, or released table). Live-pipeline intervals resample tables conditional on one frozen producer/selector run; they do not estimate run-level model variation. Local-controller overhead uses 5,000 paired resamples.

Generic early postcondition gate. An independently coded completed-return check recomputes status, answer shape, method, argument, request/document/version, and evidence bindings from raw fields before agreement features; it does not consume TEFS's saved admission IDs. It rejects the same 3,005 failed executions, 48 successful output-shape mismatches, and 1,363 out-of-policy specialist operations, with zero record-level pass-set disagreements. Its ordered view, legal scores, and selected admissible ID match TEFS on all 6,290 requests under both call-authorization settings (11 abstentions each). The wrapper fields remain trusted in both implementations; this same-system comparison tests whether the RNI view is implementation-independent under the stated trust model.

Separate-collection retrieval pilot. On a 240-instance MuSiQue-Full pilot, an existing lexical ranker and a fixed dense ranker returned 2,880 passage records for a frozen reciprocal-rank-fusion selector. A typed early check rejected zero successful returns, so on/off passage selections were identical on all 240 instances; support-passage recall@6 was 0.650 in both arms on 120 answerable instances. This null pilot probes the same selector-view property on a related *passage-selection* object. The retrievers stamped records from the same trusted corpus; support-passage recall is therefore the relevant outcome, while binding-fault incidence and answer EM are outside this pilot.

Separate-codebase answer-pool extension. On the same previously inspected 240 MuSiQue instances, pinned Llama-3-8B and Qwen3.6-27B models each answered the BM25 and dense top-six passages, creating 960 new completed records. A frozen four-answer consensus selector and typed early check were replayed over identical returns. All 960 passed the visible request/model/prompt/passage and output-shape checks; thus the on/off views, answers, and abstentions coincide on all 240 requests. On the predefined answerability-aware metric (answer EM for 120 answerable, explicit unknown for 120 unanswerable), both arms score 146/240 (60.83%), versus 151/240 (62.92%) for either Qwen single-producer control. Answerable EM is 33/120, and 113/120 unanswerable requests receive a correct unknown. The selector was constructed for this pilot. Because every completed record passes the visible contract, the experiment provides a clean null case for RNI: the selector view is unchanged and both gate arms coincide. Reference entailment is unavailable for this pilot.

Supplied typing-label diagnostic. A table-random packet contains all ten operations on each of 20 TABVERSE tables (200 requests, 1,000 invocations), without answer gold, admission verdicts, or selector outcomes. The supplied CSVs mark 21/200 typed requests as not matching the question: all 20 `partition_table` requests and one `locate_cell` request with an empty target. Eight `locate_cell` requests are flagged uncertain. The frozen `partition_table` prompt asks for first and last cell values, while the operation is excluded from the direct-extraction policy; this is a contract-semantics issue requiring adjudication, not a post hoc policy fix. Reviewer A/B exports are byte-identical, including notes; independent agreement is therefore unavailable from these exports. We use the packet as a typing-sensitivity diagnostic rather than an inter-annotator agreement study. All 1,000 invocation-conformance labels are uncertain because the packet lacks actual prompt/call evidence. A later supplement matches planned-prompt and archive-provenance hashes on all 1,000, linking the supplied labels to the planned runtime interface. Of 90 evidence-type "no" labels, all are failed returns with no evidence, not successful binding faults.

Full-ledger typing sensitivity. A retrospective deterministic scan finds that all 629 `partition_table` reader prompts request first and last cell values, whereas their specialist prompts state only the generic typed operation; 19/629 `locate_cell` requests have a blank target argument and prompt. These are structural flags, not adjudicated population error labels. On the same 6,290 frozen return sets and selector, neither flagged stratum contains an admissible-winner change; all 82 occur among the other 5,642 requests (1.45% of that stratum; 82/1,395 requests with a rejected successful return). The flagged partition requests have 23.21% answer EM under either gate arm; the 19 blank-target requests have 0% EM. The scan therefore strengthens attribution of the 82 changes to selector placement while preserving contract acquisition as a separate upstream problem.

Further scope limits. Permissive policy can admit weak alternatives, while strict policy can reject correct answers; policy-sensitivity analysis is not complete. Equation 3 does not cover generation-time shared state or count, timing, and resource side channels; read-only monitoring cannot undo producer side effects. Local mutation protection is not authenticated across services, and record provenance can itself be sensitive. The local controller-cost profile excludes pool-size scaling, serialization, contention, and throughput.

Cache-donor denominator. The donor constructor groups 629 tables by table ID and takes the next of six direct operations on the same table, producing $629 \times 6 = 3{,}774$ potential pairs. A donor must be successful and well-formed before injection. This excludes 692 failed and 48 successful but malformed source records, leaving exactly 3,034 archived control members. Because the donor constructor selects a different direct operation, an exact operation-and-arguments postcheck deter-

Table 7: All seven placement arms on the same 6,290 natural TABVERSE requests. The first five arms coincide under this frozen selector and its four-reader tie priority; the equality is empirical, not general.

Boundary	Calls	Rejected visible	Score changed	Top / choice	EM
No enforcement	31,450	1,411	1,042	274 / 82	50.40
Contract feature only	31,450	1,411	1,042	274 / 82	50.40
Method-only postcheck	31,450	1,411	1,042	274 / 82	50.40
Final mask	31,450	1,411	1,042	274 / 82	50.40
Post-score mask	31,450	1,411	1,042	274 / 82	50.40
Pre-execution authorization	28,934	48	0	0 / 0	50.30
TEFS pre-feature admission	31,450	0	0	0 / 0	50.30

ministically rejects every donor; the earlier call-time-only arm still exposes all 3,034 because the planned victim call itself is authorized. No gold answer enters donor eligibility. This is an injected substitution, not an observed cache-key collision.

Table 8: Controlled returned-record probes, all injected rather than natural incidence. Eligible cache donors number 3,034; nine stale-binding mutations hold request ID fixed.

Returned-record check	Rotated operation admitted (/6,290)	Cache donors admitted / selected (/3,034)	Held-ID stale admitted (/9)
Shape/status + final mask	4,445	3,034 / 10	–
Method-ID postcheck	3,082	3,034 / 10	–
Operation+arguments postcheck	0	0 / 0	–
Request-ID record	0	0 / 0	9 / 9
TEFS full record checks	**0**	**0 / 0**	**0 / 9**

Table 9: Leave-one-factor-out signatures from the 2^6 programmed design. Each factor has a disjoint probe stratum so other enabled checks cannot hide its omission. Only timing uses the natural replay; other rows are controlled faults, not prevalence estimates.

Disabled factor	Newly exposed endpoint	Evidence class
Pre-feature timing	1,411 illegal feature participants; 82 changed choices	natural replay
Operation check	3,082 operation-mismatched records	rotated returns
Argument check	263 misbindings; 92 answer harms	controlled point
Complete binding	9/9 held-ID stale mutations; cache donors already blocked by operation check	controlled mutation
Mutation protection	10/10 post-validation mutations	deterministic probe
Atomic emission	10/10 cross-record recombinations	deterministic probe

MMDocBench controls. On all 4,338 requests, blindly prioritizing 2,708 available structural proposals causes 127 repairs but 1,691 answer harms. The supplied-request contract admits 263 proposals, yielding 111 repairs and 17 harms. The raw pairwise arm evaluates 2,708 saved decisions; pre-admission replay reuses a decision only for an unchanged admitted pair and otherwise follows the singleton rule. Both arms share request IDs, records, and pinned model revision, but selector-call counts differ. A gold-free reason-code audit assigns 352/363 excluded pairwise choices to requests without a supported specialist contract and 11/363 to a prohibited ROI method. These choices span 14 task families and include 11 repairs versus 93 harms.

Paired live selector replay. Every completed live Candidate Record is saved once per run. In the primary replay, admission off exposes only successful rejected returns before the selector and applies a common late final mask. In each of three runs, 46/629 prompts change; the other 583 request decisions are reused because on/off inputs coincide. Both arms use the same selector implementation and, where a call is needed, the pinned Qwen2.5-VL-3B checkpoint, prompt builder, index parser, greedy settings, and vLLM backend. Eleven on-arm requests have one admitted candidate and need no model call. Complete prompt strings and raw replies are archived. Rerun on-arm winners match the original live winners 46/46 in every run. The off arm produces 17, 17, and 18 incompatible raw choices (late-masked to abstentions) plus five changed admissible winners in each run. The earlier CPU diagnostic exposed failed/empty records too: it changed 114 prompts and found 21 incompatible choices and 11 other legal- winner changes. These off-arm definitions must not be pooled. Strict answer–evidence joint accuracy is unavailable without evidence-level gold.

Nonzero-temperature sensitivity. The predeclared temperature-0.2/top-$p = 1$ study uses seeds 20260926–20260928 on the same 629 requests and the same three injected canary IDs. Each run freezes all completed returns before scoring or on/off replay. Each replay changes 46 prompts; rerun on-arm winners match the original pipeline on all 46 affected requests in every seed. The three on/off full-denominator EM pairs are 54.85/54.69, 55.17/55.17, and 54.69/54.69%. Off-arm incompatible raw choices are 18/15/8, admissible-winner changes 8/2/10, and on/off abstentions after the common late mask 0/18, 1/15, and 0/8. Natural producer failures are 122/127/121, with three additional injected canary failures per run. The on-arm mean is 54.90% with sample SD 0.24 points; off-arm mean is 54.85% with sample SD 0.28 points. These same-document seed runs quantify selector sensitivity to sampling and are reported as descriptive run-level variation.

Live binding and repeat audit. The archived cold ledger contains 1,636 admitted and 128 rejected returns: 125 natural failed or malformed outputs and three injected wrapper failures. It contains no successful, well-formed post-return binding mismatch. All 106 observed warm cache hits match the current request, operation, and document and are admitted. Forty archived repeated identical-input executions agree in prompt, raw selector reply, and selected answer; warm/cold answers agree 117/120, although five prompts differ. The two new full primary executions use the archived model revisions and greedy settings with the 1,003,520-pixel preprocessing cap; their selected answers differ from the archived run on 9 and 15 requests, respectively. These same-request repeats characterize end-to-end run stability; deployment binding prevalence is addressed separately by the live binding audit.

Interface conformance and component costs. Transfer replays admit 263/263 method-compatible MMDoc records, 110/110 completed DOGR direct-transcription records, and 3,082/3,082 TABVERSE typed records. They reject 263/263 controlled MMDoc argument-mismatch variants, 484/484 completed incompatible DOGR crops, and 1,363/1,363 incompatible TABVERSE records. In a separate 500-request profile, structural lookup plus controller costs 6.44 ms mean / 21.43 ms P95, versus 10.23 s mean / 25.32 s P95 for Qwen3.6 inference, with 57.84 GiB peak allocated model memory. The 4,338-document index occupies 1.15 GiB and takes 8.93 h to build. These measurements exclude pool-size scaling, authenticated cross-service serialization, contention, and deployment throughput.