



American Journal of Financial Technology and Innovation (AJFTI)

ISSN: 2996-0975 (ONLINE)

VOLUME 4 ISSUE 1 (2026)



PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

Version-Bounded Filing Question Answering: A Controlled Study of Source Labels, as-of Filtering, and Model Heterogeneity

Ayush Ojha^{1*}

Article Information

Received: August 01, 2026**Accepted:** October 05, 2026**Published:** October 24, 2026

Keywords

Abstention, Amendment, Financial Question Answering, Source Eligibility, Structured Output

ABSTRACT

Financial question answering must distinguish a source that contains a relevant disclosure from a source eligible at the question's cutoff. Evaluation also has to distinguish successful source selection from failures that prevent a narrow error from being counted. We examine these issues in a deterministic synthetic study of 32 original-amendment conflicts crossed with three cutoff regimes. Three prompt interfaces, four attested local model builds, and two fixed repetitions yield 2,304 retained responses. The original internally frozen cross-model hypothesis was unsupported: the mean filtered-minus-mixed wrong-version risk difference was -0.091796875 , and only two of eight strata improved. A retrospective descriptive decomposition separates response-contract failure, valid abstention, correct answering, competing-version adoption, and other responses without changing frozen scores. Two builds produced no correct conflict answers in any arm while recording zero competing-version events. Their zeros therefore cannot establish successful version selection. For qwen3:8b, correct conflict answers increased from 48/128 under mixed context to 117/128 under filtering, while counted competing-version events fell from 47 to zero. A worked response shows how a non-Boolean abstention field excludes an otherwise competing answer and citation from that endpoint. Version labels alone gave no consistent benefit. The contribution is an evaluation diagnostic for a finite synthetic suite: response validity, abstention, answer correctness, and source use must be interpreted together. The study does not establish effectiveness on real filings or general temporal reasoning.

INTRODUCTION

Financial document systems often encounter several disclosures about the same subject. An original filing can be followed by an amendment, while a user asks what was known at an earlier date. A later document may contain the most recent value in the collection and still be ineligible for that question. Conversely, an older source can remain visible after an eligible amendment has replaced the queried field. A system that finds a relevant sentence therefore still needs a rule for selecting the applicable version. This issue matters for financial technology because historical research, audit reconstruction, and source-grounded analytical workflows depend on the information boundary specified by the question.

We study a simplified boundary: one synthetic original disclosure and an amendment that replaces one categorical field. The model returns its value, source identifier, and abstention flag through a fixed JSON contract. Exact constructed answers include cases before either document exists. Retrieval, extraction, issuer interpretation, and calculation are excluded so that supplied contexts and emitted responses can be audited directly. The task is not representative of a real filing pipeline.

The original question was whether an interface that filters out future documents would reduce competing-version adoption across four selected local model builds. The frozen primary event required a schema-valid, non-abstaining response that returned both the competing value and its competing document identifier.

That event captures a specific committed error. It does not capture every unsuccessful response. An abstention on an answerable case fails the task without selecting the competing source. A response that violates the output contract also fails the task but cannot enter the event under the frozen scoring rule. Consequently, zero recorded competing-version events can mean successful source selection, refusal to answer, failure to follow the interface, or some mixture of these outcomes.

This distinction became consequential in the retained experiment. Two builds never answered a version-conflict case correctly in any arm. Both nevertheless had zero counted competing-version events. One usually failed the response contract; the other frequently emitted structurally valid abstentions even when a correct answer was available. In a different build, filtering both removed observed competing-version events and increased correct answers. These patterns cannot be represented adequately by a single error rate. Reporting the complete response states exposes what an endpoint can observe and what it leaves outside its positive class.

Recent work already documents temporal source leakage and decomposes failures in financial retrieval-augmented question answering. El Lahib et al. (2026) examine documents returned by search-engine date filters that cross a requested time boundary. Kobeissi and Langlais (2026) use document and page oracles to distinguish retrieval failures on long financial documents. These studies establish that temporal source availability and stage-level

¹ Georgia Institute of Technology, United States* Corresponding author's e-mail: aojha47@gatech.edu

diagnosis are existing research problems. Our contribution is narrower: we report how a fixed answer-and-citation error endpoint behaves when response-contract failures and abstention dominate a small version-conflict task. The study concerns generation from supplied contexts, with an exact synthetic authority rule, rather than the effectiveness of a search engine or a financial retriever.

Related financial evaluations also separate intermediate compliance from the intended outcome. Ojha (2026b) distinguishes the structural validity of financial action proposals from their effects in a controlled simulator. Ojha (2026c) reports raw field-support attempts, deterministic gate invariants, and useful completion separately. Ojha (2026a) analyzes paired response states in a synthetic payment workflow and limits claims about formatting and reasoning. These distinctions motivate the present reporting choice without supplying evidence about amendment reasoning. The task, prompts, responses, and source-selection labels in this paper are separate from those experiments; selected local builds and runtime infrastructure overlap.

We retain the original hypothesis, outcome definitions, complete response matrix, and negative support decision. We add a retrospective descriptive decomposition and an actual worked response to make the endpoint boundary visible. We do not relabel invalid responses as primary events or revise the support rule after observing the results. The resulting article contributes an auditable diagnostic case, evidence of model-by-interface heterogeneity, and a reporting approach that distinguishes correct answering from error absence. Claims are limited to 32 constructed conflicts, three interfaces, four attested builds, and two fixed repetitions. No conclusion about real issuer disclosures, legal amendment effects, investment quality, or a model population follows from this design.

LITERATURE REVIEW

Temporal source eligibility

Temporal eligibility is a property of the question and source history, not simply topical relevance. El Lahib et al. (2026) audit temporal leakage in date-filtered web retrieval. Their published study concerns the behavior of live search interfaces and returned pages. Its recommendation to use frozen snapshots is a proposed mitigation, rather than an experimentally demonstrated result for our setting. The connection is the need to distinguish a requested time boundary from the apparent relevance of retrieved text. The difference is the evidence surface: our sources have deterministic acceptance timestamps and explicit amendment relationships, while a web retrieval audit must establish dates and search behavior for actual pages.

Cymbler et al. (2026) provide a close temporal-grounding precedent in a versioned French tax-law corpus. They distinguish strict answer measures from retrieval-side provenance applicable at the query date. Their work establishes that answer and date-applicable source evidence can be measured separately in a real retrieval setting. Huwiler et al. (2025) study version and change

retrieval over technical documentation; their stated limitations distinguish discrete release versions from calendar-time-to-release reasoning. We cite both as preprints and do not treat their findings as validation of our financial setup. The new contribution is consequently the response-contract and abstention diagnostic under the constructed filing rule, rather than the general concept of version-sensitive grounding.

Financial question answering and failure decomposition

Kobeissi and Langlais (2026) study retrieval failures in long-document financial question answering using FinanceBench questions and document/page oracles. Their oracle comparisons distinguish failures that arise when evidence must be located from behavior when the relevant document or page is provided. This is relevant because an aggregate answer score can combine errors from several stages. Our evaluation begins after context selection, so it cannot attribute failures to retrieval. It extends the diagnostic question to the emitted response: even after the source package is fixed, a narrow wrong-source event can fail to expose contract violations and refusal to answer.

An oracle-provided page and a timestamp-filtered synthetic source package are not interchangeable controls. A page oracle addresses where evidence appears in a long document. Our authority labels address which constructed version is eligible at a cutoff. Neither setup alone establishes amendment extraction, historical source completeness, or reliable financial reasoning. The comparison supports decomposition as an evaluation method rather than a claim that the same benchmark or mechanism has been replicated. A future real-data study would need both stage-specific retrieval diagnostics and independently reviewed source-eligibility labels.

The present use of categorical values also avoids the arithmetic and interpretation burden of many financial tasks. A correct response selects a literal state such as limited or heightened and the exact synthetic document identifier. There is no estimate to calculate, unit conversion to infer, or materiality judgment to make. Exact scoring is defensible for these labels because the expected strings are defined by construction. It should not be generalized to open-ended financial answers, where synonyms, incomplete evidence, and multiple defensible explanations can require a different evaluation protocol.

Structured responses and useful completion

Ojha (2026c) evaluates typed field support with a value-only deterministic executor on synthetic financial records. The paper distinguishes candidate attempts from outputs that pass a gate and from successful completion. Its post-gate invariants are conditional on what the executor permits; they do not establish that every raw attempt was reliable. This distinction is directly relevant to the current primary endpoint. A strict response validator can prevent an invalid response from entering a semantic category

while leaving the task unresolved. The financial-record evaluation concerns typed support and deterministic calculation, whereas this paper concerns source versions and cutoff-specific authority.

Ojha (2026b) similarly separates valid proposal structure from harm and utility in a financial action simulator. The model's proposal and a later state transition are different objects of evaluation. Structural acceptance does not establish that an action satisfies the intended outcome. Here, the separation runs in both directions: schema compliance is necessary for scoring a committed competing-version response, but failure to trigger that event does not establish correct answering. Both studies motivate reporting the intermediate condition alongside the final task outcome. Neither provides evidence that source filtering improves filing question answering.

Ojha (2026a) examines presentation order and structured diagnosis in a synthetic payment workflow. Its paired response-state analysis preserves fixed denominators and distinguishes valid, invalid, and task-relevant response behavior. It also avoids treating a formatting change as a demonstrated reasoning mechanism. The present retrospective categories follow that reporting principle: they are a descriptive account of retained responses under an unchanged denominator. They are not a new confirmatory endpoint, a human semantic adjudication, or an independent replication of the payment benchmark.

Context presentation and attribution limits

Liu et al. (2024) show that language-model use of long

contexts can depend on where relevant information is placed. That work makes position an important design consideration, but it does not explain the results of our short two-document prompts. We alternate document order deterministically by base-conflict parity. This reduces perfect alignment between source role and position without randomizing positions or estimating a positional effect. Our arms also differ in labels and document exposure. A result for an implemented interface can therefore reflect several changes at once.

Our contribution is the reconciled response matrix and endpoint boundary under an exact constructed version rule. Temporal retrieval, oracle-based financial QA diagnosis, and the separation of intermediate acceptance from useful outcomes already have precedents. We make no priority claim for these broader ideas.

MATERIALS AND METHODS

Design and evidence boundary

The experiment follows the internally frozen protocol VBQA-SYNTH-1.0. Thirty-two base conflicts are crossed with three cutoff regimes, producing 96 logical cases. Every case is rendered through three source-policy arms and presented to four selected local model builds in two fixed repetitions. The full matrix contains $96 \times 3 \times 4 \times 2 = 2,304$ responses. All comparisons are matched within case, model, and repetition. Table 1 lists the design dimensions. The count follows from a complete enumerated matrix; it is not a sample-size estimate for a population study.

Table 1: Design dimensions of the retained experiment

Dimension	Levels	Count
Base conflict	Eight fields with four categorical transitions each	32
Cutoff regime	Pre-filing, between versions, post-amendment	3
Logical case	Base conflict crossed with cutoff regime	96
Source-policy arm	Mixed, labeled, filtered	3
Attested local build	Four operator-selected installed builds	4
Repetition	Fixed seeds 41873 and 41874	2
Retained response	Complete Cartesian product	2,304

The protocol and implementation were bound to an August 30, 2026 local freeze manifest before the eligible execution. The freeze records the design, prompt arms, scoring rules, repetitions, and support criteria. This is an internal protocol freeze, not externally registered preregistration. Hashes establish consistency with a recorded local file set; they do not independently establish authorship or a trusted public chronology. The authoritative execution realizes the complete planned matrix. Two later incomplete duplicate attempts were preserved as quarantined incidents and excluded from scoring, replacements, sensitivity analyses, and the retrospective decomposition. Supplement S1 records their identities and the unchanged evidence bindings.

The four builds were selected from locally available

installations by the operator. They were not drawn from a model population. The 32 conflicts enumerate the generator's templates rather than a sample of issuers or filings. Both features limit generalization. Repetition captures two retained executions under the fixed protocol; it does not make the constructed documents independent or establish deployment reliability. Our main analysis reports finite-suite counts and matched descriptive differences. Historical statistical calculations remain available in S1 with their calibration limitations.

Documents and authority labels

The generator defines eight categorical disclosure fields: risk outlook, supply strategy, product availability, audit status, legal proceeding status, debt covenant status,

cyber incident status, and revenue-recognition basis. Each has four fixed transitions between literal values, giving 32 original–amendment pairs. The labels are artificial states. They were not extracted from issuer documents or paraphrased from a financial dataset. Each pair has a synthetic issuer name, an original identifier ending D01, and an amendment identifier ending D02.

The original acceptance timestamp is January 15, 2040, at 12:00 UTC. The amendment acceptance timestamp is February 15, 2040, at 12:00 UTC. These future dates are fixed experimental symbols, not actual filing history. Each source includes its identifier, issuer, timestamp, queried field, value, and a sentence identifying it as deterministic synthetic material. Labeled sources additionally state their original or amendment role. The amendment identifies its target and says that it replaces only the queried field. This field-level relationship prevents the constructed oracle from treating an amendment as a replacement for unrelated information.

Three cutoffs define the regimes. At January 1, 2040, neither source is eligible, so the correct response is abstention with null answer and citation. At February 1, 2040, only the original is eligible and its value is authoritative. The amendment is the competing version. At March 1, 2040, both are eligible and the amendment is authoritative for the queried field. The original is the competing version. The latter two regimes form the 64 logical conflict cases used by the primary endpoint. Pre-filing cases have no eligible authority and no competing answer under that endpoint.

The generator stores the expected value, expected document identifier, expected abstention flag, competing value, competing identifier, and eligible identifier list outside the response schema. These fields define the exact oracle. Original–amendment order alternates by base-pair parity: odd pairs present original first and even pairs reverse the order. The same order rule is applied to labeled and unlabeled contexts and to the eligible subset. There is no outcome-dependent case generation, document selection, or authority adjudication.

Implemented prompt interfaces

All prompts tell the model to use only sources accepted on or before the cutoff, apply an amendment only to the field it replaces, use the latest eligible disclosure when labels are absent, and abstain when no source is eligible. The question, categorical values, timestamps, and three-field response contract are fixed across arms. The mixed interface presents both sources at every cutoff without version-role or replacement-scope lines. Timestamps remain visible. The model must infer the applicable version from the supplied timing and instruction.

The labeled interface also presents both sources at every cutoff but adds the original/amendment roles, amendment target, and field-level replacement statement. A future amendment therefore remains visible in this arm when the cutoff makes it ineligible. The filtered interface retains those labels and removes sources accepted after

the cutoff. Its pre-filing context contains no document, its between-version context contains only the original, and its post-amendment context contains both sources. Filtering enforces the synthetic eligibility rule before generation.

Mixed-to-filtered comparisons change metadata and document exposure together. They estimate the descriptive difference between implemented packages, not an isolated effect of temporal reasoning. Labeled-to-filtered comparisons are also regime-dependent. After the amendment, their prompt contexts are identical because both documents are eligible. Differences between their retained outputs in that regime cannot be attributed to removal of an ineligible document. Shorter context, reduced contradiction, instruction burden, and execution variation are possible contributors elsewhere; the design does not distinguish these explanations.

Execution and model identity

Each case-arm prompt was executed twice with temperature zero and seeds 41873 and 41874 through the local Ollama generate endpoint. Requests were nonstreaming and requested JSON format. The selected builds were qwen3:8b, gemma3:4b, llama3.1:8b, and phi4-mini:3.8b, attested under Ollama 0.32.14. Names identify the observed installations rather than an abstract model family. Full recorded digests appear in S1. Model choice in executable code is environment-driven; no build was substituted after observing an outcome.

Before execution, the runner checked the frozen file set and canonical generated cases, reserved new raw and sidecar files exclusively, and recorded the installed model inventory and runtime version. Every raw response row binds a protocol identifier, generated-case hash, runtime-attestation hash, case, arm, build, repetition, seed, textual response, and provider metadata. The runner flushes retained rows to an append-only journal. An interruption or provider failure leaves an incomplete file that cannot satisfy the complete-matrix analysis gate. Existing raw files are not overwritten or silently resumed.

The eligible evidence has 2,304 unique case-arm-build-repetition keys and no missing matrix cell. For this revision, a separate strict parser independently reconstructed the oracle from generator fields and checked every retained response against the frozen score indicators. It also reconciled all twelve build-arm aggregate groups. There were zero mismatches. This is computational verification of records and scoring, not independent execution of the model experiment or human peer review. No new model calls were made for the manuscript revision.

Frozen response contract and endpoints

The response must be exactly one JSON object with exactly three keys: answer, cited_document_id, and abstain. Answer and citation must be strings or JSON null; abstain must be Boolean. Extra or missing keys, duplicate keys, wrappers, invalid syntax, wrong types, and other contract violations are failures. We call this combined

category response-contract failure. The historical scorer names it parse failure, but not every instance is syntactically invalid JSON. A response with abstain:null can parse as JSON and still violate the Boolean schema. Scoring uses exact case-sensitive equality without synonym mapping, whitespace rescue, fuzzy matching, a model judge, or post-hoc human normalization. Joint correctness requires the expected answer, citation, and abstention flag together. A pre-filing correct response is a valid abstention with null answer and citation. A conflict-case correct response is a non-abstaining answer with the authoritative value and source. These are both joint-correct under the frozen definition, but they represent different task outcomes and are separated in the new descriptive reporting.

The primary competing-version-adoption event requires a structurally valid non-abstaining response that returns both the competing value and competing document identifier. The frozen wrong-version answer and citation indicators are its components. A contract failure receives false on all correctness indicators and cannot enter the primary event. An incorrect abstention similarly cannot enter it. The frozen unavailable-source indicator separately records a valid response citing an identifier outside the cutoff-eligible list. These rules are retained exactly; the revision changes their presentation rather than their classification.

Retrospective complete-outcome decomposition

We partition each retained response into five mutually exclusive descriptive states. F is response-contract failure. A is a structurally valid abstention. C is a jointly correct non-abstaining answer. W is the frozen competing answer-and-citation event. O is any remaining valid non-abstaining response. Classification checks contract validity first, then abstention, then correct answering, then the primary event. C and W cannot overlap because every constructed original-amendment pair has different values and identifiers. The five counts therefore sum to the full retained denominator.

All observed A responses had null answer and citation. A is successful only in the pre-filing regime; on a conflict case it is an incorrect refusal. Frozen joint correctness J equals C plus correct pre-filing A in these records. W's internally frozen denominator remains 128 conflict responses per build-arm group, while F/A/C/W/O use all 192 responses. Each regime-specific cell contains 64 responses, from 32 base conflicts and two repetitions. We keep these denominators visible rather than calculating accuracy only among valid or answering responses.

This partition was added after inspecting the completed evidence. It is retrospective and descriptive, not preregistered or a replacement confirmatory hypothesis. It does not infer the semantic intention of an invalid response. F includes outputs that may contain the competing value, but those responses remain F and retain W=false. The worked case illustrates this boundary using an actual row. It is not a qualitative rescue procedure

applied to the matrix.

Original support rule and interpretation of uncertainty

The original H1 requires all eight filtered-minus-mixed wrong-version risk differences to be nonpositive, their mean to be at most -0.10 , at least six to be strictly negative, and at least four Holm-adjusted exact McNemar p-values below .05. The protocol also bars a simple improvement claim if any arm-level joint-accuracy loss exceeds .02 in a model-repetition stratum. The unchanged analyzer's claim flag requires H1 support and checks accuracy loss in the primary contrasts; it is false. We retain the negative decision and disclose the labeled-arm accuracy loss rather than treating the primary flag as a check of every secondary safety tradeoff.

The frozen analyzer also reports Wilson intervals and binomial-reference McNemar calculations. Each of 32 base conflicts contributes two cutoff-specific observations sharing constructed documents and a question. The design supplies neither a probability sample nor a validated independence or exchangeability argument for that reference law. Repeated model executions do not resolve the shared-case dependence. Consequently, main-text comparisons use counts and descriptive paired differences. S1 preserves the historical intervals, tests, adjustment, and exact support-rule values with an explicit statement that population uncertainty and clustered calibration were not established.

RESULTS AND DISCUSSION

Complete response states

Table 2 reports every build-arm group using the five-state partition. Each row sums to 192 responses. Frozen joint correctness is shown separately because it combines correct abstention and correct answering. This makes the distinction visible without changing the internally frozen endpoint. Three builds record zero W in every arm, but their C counts show that this absence has different implications from the high correct-answer count in filtered qwen.

Gemma produced 510 contract failures across 576 responses. Its filtered arm improved J from zero to 64, entirely through correct pre-filing abstentions. Both conflict regimes still had 64/64 failures in that arm. Phi produced 544 failures across 576 responses. Its mixed arm contained one correct post-amendment answer; its filtered arm contained 31 correct pre-filing abstentions and no correct conflict answer. For these builds, filtering changed the count of successful empty-context responses while leaving the source-selection task largely unperformed.

Llama illustrates a different route to zero primary errors. Its mixed arm produced 190 valid abstentions and two contract failures. Sixty-two abstentions were correct before filing; the 128 conflict abstentions were incorrect refusals. Its labeled arm produced 100 contract failures and 92 abstentions, while its filtered arm produced 46

Table 2: Complete response states and frozen joint correctness by build and interface

Build	Arm	F	A	C	W	O	J
gemma3:4b	Mixed	190	0	0	0	2	0
gemma3:4b	Labeled	192	0	0	0	0	0
gemma3:4b	Filtered	128	64	0	0	0	64
llama3.1:8b	Mixed	2	190	0	0	0	62
llama3.1:8b	Labeled	100	92	0	0	0	30
llama3.1:8b	Filtered	46	146	0	0	0	64
phi4-mini:3.8b	Mixed	191	0	1	0	0	1
phi4-mini:3.8b	Labeled	192	0	0	0	0	0
phi4-mini:3.8b	Filtered	161	31	0	0	0	31
qwen3:8b	Mixed	46	0	48	47	51	48
qwen3:8b	Labeled	42	0	49	49	52	49
qwen3:8b	Filtered	11	64	117	0	0	181

Note. F = response-contract failure; A = valid abstention; C = correct non-abstaining answer; W = competing answer and citation; O = other valid non-abstaining response. J is the frozen joint-correct count. F/A/C/W/O each use 192 as their group denominator; the internally frozen W rate uses 128 conflict responses. J is not a sixth partition state.

failures and 146 abstentions. It never produced a correct non-abstaining answer in any arm. The increase in J from 62 to 64 under filtering therefore reflects two additional correct pre-filing abstentions rather than improved answering of an eligible disclosure.

Qwen generated enough valid committed answers for W to be informative about competing-version use. Mixed context produced 48 correct conflict answers, 47 W events, and 46 contract failures across all regimes. Filtered context produced 117 correct conflict answers, 64 correct pre-filing abstentions, eleven failures, and no W events. Thus the rise in J from 48 to 181 combines 69 additional correct conflict answers with 64 correct abstentions. Both components matter, and neither should be hidden inside the aggregate improvement.

Cutoff-specific diagnosis

Table 3 separates the regimes. Each entry lists F/A/C/W/O and sums to 64. The pre-filing regime tests recognition that no eligible disclosure exists. The between-version regime tests resistance to an explicitly future source. The post-amendment regime tests use of the amended field once both sources are eligible. This decomposition was descriptive, but the regime definitions themselves were frozen before execution.

All 47 mixed qwen W events occurred between versions, when the future amendment should not have been used. The mixed post-amendment arm recorded no W event and 46 correct answers, with eighteen contract failures. Filtering yielded 64/64 correct answers between versions and 53/64 after amendment. This pattern is compatible

Table 3: Response states by cutoff regime with a fixed 64-response denominator per cell

Build	Arm	Pre-filing	Between versions	Post-amendment
gemma3:4b	Mixed	62/0/0/0/2	64/0/0/0/0	64/0/0/0/0
gemma3:4b	Labeled	64/0/0/0/0	64/0/0/0/0	64/0/0/0/0
gemma3:4b	Filtered	0/64/0/0/0	64/0/0/0/0	64/0/0/0/0
llama3.1:8b	Mixed	2/62/0/0/0	0/64/0/0/0	0/64/0/0/0
llama3.1:8b	Labeled	34/30/0/0/0	34/30/0/0/0	32/32/0/0/0
llama3.1:8b	Filtered	0/64/0/0/0	14/50/0/0/0	32/32/0/0/0
phi4-mini:3.8b	Mixed	64/0/0/0/0	64/0/0/0/0	63/0/1/0/0
phi4-mini:3.8b	Labeled	64/0/0/0/0	64/0/0/0/0	64/0/0/0/0
phi4-mini:3.8b	Filtered	33/31/0/0/0	64/0/0/0/0	64/0/0/0/0
qwen3:8b	Mixed	13/0/0/0/51	15/0/2/47/0	18/0/46/0/0
qwen3:8b	Labeled	12/0/0/0/52	15/0/0/49/0	15/0/49/0/0
qwen3:8b	Filtered	0/64/0/0/0	0/0/64/0/0	11/0/53/0/0

Note. Entry order is F/A/C/W/O as defined in Table 2. Abstention is correct only in the pre-filing column. No regime cell is a new independent sample of 64 base conflicts.

with preferring the latest visible version: the amendment is ineligible between versions but authoritative afterward. The counts locate the observed errors; they do not distinguish that heuristic from other explanations or isolate which feature of the filtered package caused the change.

The labeled and filtered post-amendment contexts are identical. Their retained qwen outcomes nevertheless differ: 49 correct and fifteen failures for labeled, compared with 53 correct and eleven failures for filtered. Those four additional correct responses cannot be interpreted as evidence of document filtering in that regime. The experiment did not establish bitwise deterministic inference, and the arm executions were separate requests. The equal-context comparison cautions against attributing every between-arm difference to source removal.

Pre-filing behavior explains much of the apparent

aggregate benefit outside qwen. Filtered prompts contain no document in that regime. Gemma answered these with 64 correct abstentions; phi with 31; llama with 64. These are useful successful responses to a no-evidence question, but they do not demonstrate choosing the correct version when evidence exists. A financial QA report that combines abstention and answering should therefore publish their separate counts whenever the task includes both.

A recorded response that exposes the endpoint boundary

Consider SYN-0001-BETWEEN_VERSIONS in qwen repetition zero. The queried field is risk outlook. Original SYN-0001-D01 discloses limited and is accepted January 15, 2040, at 12:00 UTC. Amendment SYN-0001-D02 discloses heightened and is accepted February 15, 2040,

Table 4: Actual responses for one between-version case in qwen repetition zero

Arm	Answer	Citation	Abstain field	Frozen classification
Mixed	heightened	SYN-0001-D02	null	Contract failure; W=false; J=false
Labeled	heightened	SYN-0001-D02	false	Competing-version event; W=true; J=false
Filtered	limited	SYN-0001-D01	false	Correct answer; W=false; J=true

at 12:00 UTC. The cutoff is February 1, 2040, at 00:00 UTC. The original is therefore the only eligible source; limited/SYN-0001-D01 is the constructed correct answer/source pair. Table 4 shows the literal recorded fields across arms.

The mixed response is syntactically valid JSON. Its abstain:null violates the Boolean contract, so the strict parser rejects it and the primary event remains false. The labeled response contains the same competing value and citation with abstain:false, making it a counted W event. This difference shows why W=false must not be read as absence of competing-version content. It also shows why the new decomposition retains the failure category rather than silently changing the frozen labels.

The example is one recorded matched case, not a selected demonstration of population efficacy. It supports the scoring-boundary argument through literal fields and a reproducible rule. It cannot establish how many other invalid outputs intended to adopt a wrong version, because we did not conduct semantic adjudication of failures. The complete F count remains visible so that the reader can judge the extent of unclassified task failures without an invented estimate of their semantic content.

The original hypothesis remains unsupported

All eight filtered-minus-mixed wrong-version risk differences were nonpositive. Six were zero. The two qwen differences were $-22/64$ and $-25/64$, yielding the unchanged arithmetic mean -0.091796875 . This misses the frozen requirement of at most -0.10 . Only two strata were strictly negative, compared with six required. Only those two also met the internally frozen Holm

threshold, compared with four required. H1 therefore remains unsupported. Its failure follows from the mean and negative-stratum conditions even without relying on calibrated p-values.

Joint correctness increased in all eight primary strata, but this does not authorize a general improvement statement under the protocol. The build-level breakdown shows why: the increases can come from correct abstentions rather than answering. The original support rule also requires H1 first. We preserve that restriction and do not replace the failed robustness claim with an outcome-dependent success criterion. The diagnostic contribution is a new interpretation of the retained evidence, explicitly separated from confirmatory support.

Version labels alone provide no consistent benefit. Qwen W rose from $47/128$ to $49/128$, while J changed from $48/192$ to $49/192$. Llama J fell from $62/192$ to $30/192$ and contract failures rose from $2/192$ to $100/192$. Gemma and phi reached $192/192$ contract failures in the labeled arm. These results concern this particular metadata presentation and selected builds. They do not establish that version metadata is generally ineffective, but they show that correct labels placed inside a prompt are insufficient as a universal control in this experiment.

Implications for financial technology evaluation

The practical evaluation lesson is to publish the path by which a system avoids a counted error. A low wrong-source event rate is useful only alongside evidence that the system returns valid, correct answers when answers are available and abstains appropriately when they are absent. Contract failure, correct abstention, incorrect

refusal, correct answering, and other answer errors have different implications for a user. Maintaining a fixed total denominator prevents a system that rarely answers from appearing reliable merely because conditional accuracy excludes failures.

This diagnostic extends the stage-decomposition perspective of Kobeissi and Langlais (2026) to post-context responses. It is also consistent with the separation between raw attempts, gated outputs, and useful outcomes in Ojha (2026c), and between proposal validity and realized effects in Ojha (2026b). The present evidence adds a concrete case in which a conjunction endpoint and strict schema validation jointly leave a major class of unsuccessful responses outside the counted semantic error. The payment-workflow results of Ojha (2026a) support reporting paired response states while retaining caution about reasoning attribution.

For an operational filing system, external source eligibility and response validation address different responsibilities. A source filter can prevent a future document from entering context if timestamps and amendment scope are correct. A validator can ensure that the emitted structure obeys an interface. Neither by itself guarantees an answer is correct, that the eligible collection is complete, or that abstention is appropriate. Our study motivates testing these responsibilities together. It does not validate any deployed filter, parser, retriever, or financial decision process.

The strongest positive observation is confined to qwen's retained conflict responses. Its correct-answer count and W count improve together under the filtered package. That is a more informative observation than error reduction alone, but still a finite-suite result for one attested build. It cannot support a model-general claim or isolate a temporal reasoning mechanism. The negative cases are part of the evidence rather than obstacles to remove: they show when a primary event ceases to be an adequate summary of task behavior.

Limitations and next empirical test

The synthetic oracle removes ambiguity at the cost of realism. Real filings contain long prose, tables, exhibits, incorporated material, multiple revisions, and incomplete or ambiguous amendment relationships. No actual issuer, SEC accession, filing text, market price, or investor decision was used. The categorical values and two-source contexts do not model extraction, numerical reasoning, retrieval ranking, or financial interpretation. External validity is unknown. The field replacement rule is a constructed assumption rather than a legal conclusion.

The four builds are a convenience set with fixed installed artifacts. Two repetitions and temperature zero do not establish deterministic generation across hardware, request scheduling, or runtime versions. The identical post-amendment labeled/filtered prompts with different retained outputs make that limit observable. The study cannot rank model families generally or infer the behavior of other sizes, quantizations, prompts, or deployments.

Shared builds with companion studies further mean that those publications are separate tasks, not independent samples of a model population.

The endpoint is intentionally narrow and the response contract is strict. Contract failures cannot register primary adoption even if they contain a competing value. Valid abstention can avoid W while failing an answerable case. The retrospective partition makes these limitations visible but does not remove them. It contains no semantic assessment of invalid responses, no relaxed-schema rescoring, and no human-rater evidence. A study of underlying source beliefs would require a new endpoint and a new protocol rather than a post-hoc reinterpretation of the present flags.

Interface contrasts also contain confounding. Mixed-to-filtered changes relationship labels and eligible-source exposure; source removal changes context length and contradiction. Alternating order reduces one deterministic alignment but does not identify a positional mechanism. Equal-context arm differences show that separate retained executions can differ without a source-policy change. A follow-up experiment should establish baseline schema competence, vary filtering and metadata separately, and include repeated equal-prompt controls before interpreting differences as version-reasoning effects.

Finally, the case matrix is finite and dependent. Multiple regimes share a base conflict, and repetitions revisit the same constructed material. The protocol does not validate sampling uncertainty or clustered reference-law calibration. The original freeze is local and the partial duplicate incident reveals the need for stronger process isolation in future execution. A real-data follow-up would require a newly frozen protocol, public accession provenance, an independently checked amendment-scope oracle, explicit retrieval and extraction stages, and separate reporting of invalid output, inappropriate abstention, successful answering, and wrong-source use.

CONCLUSION

The retained 2,304-response experiment does not support its original cross-model filtering hypothesis. It does support a bounded evaluation diagnostic: absence of a narrow competing-version event cannot establish successful source selection when contract failure and abstention are common. Two builds produced no correct conflict answers in any arm while recording zero such events. Qwen showed a different finite-suite pattern, with 48/128 correct conflict answers under mixed context and 117/128 under filtering, alongside 47 versus zero counted competing-version events. Version labels alone had no consistent benefit.

The revision preserves the complete evidence and frozen negative decision, and adds a retrospective response-state decomposition with an actual worked case. Financial QA evaluations should make validity, abstention, correct answering, and source use jointly visible under fixed denominators. These observations concern constructed

short contexts and selected local builds. Claims about real filings, model-general temporal reasoning, or deployment reliability require a new empirical study.

Data and code availability

The generator, protocol, scoring code, complete retained responses, runtime attestation, frozen analysis, and reconciliation records are preserved locally. The complete confidential VBQA-SYNTH-1.0 reproduction archive was sent privately to the journal by email on October 12, 2026. Editor acknowledgment and acceptance of the archive remain pending. The author has confirmed the required submission rights and authorized confidential journal review. Supplement S1 documents the methods and historical calculations. No public repository, persistent identifier, or redistribution license is asserted. The experiment used synthetic documents and no human participants, identifiable customer data, or real issuer records.

REFERENCES

- Cymbler, R., Guez, D., & Fabre, L. (2026). *Temporal misgrounding in legal RAG: A versioned-corpus benchmark for French tax law* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2608.09393>
- El Lahib, A., Xia, Y.-J., Li, Z., Wang, Y., & Pi, X. (2026). Temporal leakage in search-engine date-filtered web retrieval: A retrospective forecasting case study. In M. Liakata, V. P. Moreira, J. Zhang, & D. Jurgens (Eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 787–795). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.acl-short.65>
- Huwiler, D., Stockinger, K., & Fürst, J. (2025). VersionRAG: *Version-aware retrieval-augmented generation for evolving documents* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2510.08109>
- Kobeissi, A., & Langlais, P. (2026). *Decomposing retrieval failures in RAG for long-document financial question answering* [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2602.17981>
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. https://doi.org/10.1162/tacl_a_00638
- Ojha, A. (2026a). Presentation order and structured diagnosis in a synthetic payment workflow benchmark. *American Journal of Innovation in Science and Engineering*, 5(3), 19–30. <https://doi.org/10.54536/ajise.v5i3.8542>
- Ojha, A. (2026b). Stateful mediation and selective auditing for financial language-model actions: A paired controlled simulator study. *American Journal of Financial Technology and Innovation*, 4(1), 198–210. <https://doi.org/10.54536/ajfti.v4i1.8543>
- Ojha, A. (2026c). Typed field support with a value-only deterministic executor for synthetic financial records: A multi-seed controlled evaluation. *American Journal of Financial Technology and Innovation*, 4(1), 185–197. <https://doi.org/10.54536/ajfti.v4i1.8551>