The Bias Tax: From Closure Failure to Verification Overhead in Long-Context LLM Auditing

Junzhe Cai
Independent Researcher, High School Division

March 2026

Abstract

As long-context Large Language Model (LLM) evaluation shifts from simple retrieval toward audit-like reasoning, the critical challenge is no longer merely finding relevant facts, but preserving a correct logical closure under prompt-induced pressure and understanding how generation failures propagate into downstream verification cost. We study this pipeline effect in a single 80,000-token legislative corpus using six prompting conditions—Control, Management, Chain-of-Thought (CoT), Periodic Summary, Union, and One-Shot analogy prompting—within a Reader–Judge architecture, with DeepSeek-V3 as the solver and DeepSeek-R1 as the auditor. We introduce a **Logical Needle-in-a-Haystack (L-NIHS)** stress test in which success requires traversing a five-needle chain from base rule to factual evidence and final closure while resisting a late-stage distractor. Our main result is structural: the dominant solver-side failure is not early retrieval loss, but late-stage closure failure under behavioral pressure. Across prompting conditions, early anchors remain largely recoverable, while interference rejection and closure degrade sharply once persona-congruent or prompt-congruent distractors enter the reasoning path. On the auditor side, these distorted outputs induce a shorter-but-costlier verification asymmetry—the **Bias Tax**—in which management-conditioned responses are briefer yet slower to verify than neutral ones. Judge-side reasoning traces further suggest that this added cost arises not from answer length alone, but from the need to disentangle factual closure from fabricated procedural or compensatory pathways. Finally, we show that high tone certainty is widespread across conditions and therefore cannot be treated as a reliable proxy for logical fidelity. Taken together, the results suggest that in long-context auditing, prompting strategies reshape not only answer style, but the pathway by which evidence is selected, justified, and transformed into both final decisions and downstream audit burden.

1 Introduction

Long-context LLMs are often evaluated by asking whether they can recover information buried in large inputs. That question remains important, but it is no longer sufficient. In real auditing tasks, the model must not only retrieve dispersed evidence; it must also preserve the correct infer-

ential path when the prompt pushes it toward a particular stance, interest, or narrative. The key practical question is therefore not just *Can the model find the evidence?*, but *What happens to the final closure once prompt framing starts competing with the document's factual structure?*

This problem sits at the intersection of two literatures. On one side, long-context work has highlighted positional sensitivity and multi-needle retrieval difficulty in large contexts [9, 5]. On the other side, bias and prompting research has shown that LLMs are sensitive to persona induction, non-neutral framing, and unfaithful reasoning traces [2, 1, 11, 4, 12]. Yet these lines are usually studied separately: long-context benchmarks often emphasize retrieval capacity, while bias studies often focus on style, stance, or short-context reasoning. What remains under-explored is their interaction under audit-like conditions where a model must back-trace across a long document, reject an adversarial distractor, and produce a numerically precise closure.

This paper addresses that gap with a controlled Reader–Judge study over an 80,000-token legislative corpus. The task is deliberately narrow but revealing: the correct settlement value can only be obtained by chaining a base rule, a trigger rule, a buried factual reading, and a final closure while resisting a late administrative distractor. We refer to this setup as a **Logical Needle-in-a-Haystack (L-NIHS)** stress test. Rather than claiming direct access to hidden internal states, we quantify **behavioral manifestations** of prompt pressure: answer-family divergence, needle-level retrieval and closure scores, evidence-lean direction, output length, generation latency, throughput, audit latency, and logged failure episodes.

The core claim of this paper is strong but specific. We show that the principal bottleneck in this task is **not** early retrieval. Models often recover the early anchors. The real fracture occurs later, when prompt framing changes which evidence path is privileged and whether the model successfully closes the chain. This late-stage distortion appears in answer families, interference rejection scores, closure scores, qualitative failure logs, and audit cost. We call one particularly striking efficiency pattern the **Bias Tax**: some prompt-conditioned outputs are shorter, yet still cost more to verify.

The paper therefore preserves a forceful thesis: explicit prompting strategies and persona induction do not merely change style; they can reshape logical closure in measurable ways. At the same time, we deliberately distinguish what the current evidence supports behaviorally from what would require trace-level mechanism work to prove.

2 Related Work

2.1 Long-Context Retrieval and Position Sensitivity

Long-context evaluation has moved beyond perplexity toward recovery of relevant information under extended input length. Lost in the Middle showed that retrieval quality depends strongly on position within the context window [9], while RULER extended long-context testing to more complex, multi-needle scenarios [5]. Our work is complementary to this line. Rather than asking only whether a model can retrieve multiple buried facts, we ask what happens after retrieval—specifically, whether the model maintains or abandons the correct inferential path when prompting injects behavioral pressure.

2.2 Bias, Persona Induction, and Unfaithful Reasoning

LLMs inherit biases and perspective imbalances from training data [2, 1, 11]. Persona induction can amplify these tendencies by steering outputs toward a chosen social or institutional viewpoint [4]. At the same time, reasoning traces are not always faithful to the actual basis of the answer [12, 8]. This motivates our focus on *closure fidelity*: a model may cite true facts, but still close on the wrong path after a persona-congruent or prompt-congruent distractor is admitted into the chain.

2.3 Prompting Strategies Under Pressure

CoT prompting, in-context exemplars, and decomposition strategies can improve performance in many tasks [14, 3, 10]. However, their benefits are not uniform. Prior work documents hallucination, brittle planning, and the limits of self-correction under more difficult conditions [13, 7, 6]. We extend this discussion into a long-context audit setting by comparing six prompting strategies inside the same corpus and under the same judge rubric.

3 Methodology

3.1 Task Setting: Logical Needle-in-a-Haystack

We construct a **Logical Needle-in-a-Haystack (L-NIHS)** stress test over an 80,000-token legislative corpus, *Afar Sector Labor and Environment Settlement Integrated Management Regulations*. The target answer requires a five-needle closure

$$N1 \rightarrow N2 \rightarrow N3 \rightarrow N5,$$

while resisting a late-stage distractor $N4$.

L-NIHS should be understood not as a claim of inventing needle-style long-context evaluation itself, but as an extension of that family from retrieval-oriented stress testing to **closure-oriented audit evaluation**. The goal is not merely to locate buried facts, but to preserve the correct path from rule to evidence to final numerical closure.

The functional needles are:

- **N1 (Base):** Article 8 specifies the base settlement $S_{base} = 500$.
- **N2 (Rule):** Article 27 specifies the 1.3 coefficient when ion level $< 0.05\%$ and resonance frequency $= 14.22$Hz.
- **N3 (Evidence):** Article 280 provides the relevant factual readout: ion $= 0.01\%$ and frequency $= 14.22$Hz.
- **N4 (Late-stage interference source):** administrative and sulfur-concentration distractors, especially Article 350 and Article 82.
- **N5 (Closure):** the correct synthesis $500 \times 1.3 = 650$, while rejecting the distractor pathway.

3.2 Reader–Judge Pipeline

The experiment uses a two-stage Reader–Judge architecture [15]. DeepSeek-V3 serves as the **Solver**, reading the full prompt and generating the audit answer. DeepSeek-R1 serves as the **Auditor**, scoring the Solver's response against a structured gold rubric rather than rereading the full 80,000-token corpus.

The Auditor prompt explicitly includes the five gold needles, per-needle scoring criteria, hallucination and tone scales, and extracted fields such

Table 1: Prompting conditions used in the experiment.

Group	Strategy	Description
Group 1	**Control**	Neutral instruction: compute the final settlement value objectively from the full act.
Group 2	**Management**	Strong cost-control persona: reject any non-mandatory increase and minimize payout wherever the document permits.
Group 3	**CoT**	Structured audit workflow: retrieve rules, fetch records, validate logic, then compute the final value step by step.
Group 4	**Periodic Summary**	Chunk-wise reading with intermediate summaries before final synthesis.
Group 5	**Union**	Strong labor-advocate persona: maximize legitimate compensation by foregrounding pro-worker clauses and harsh-condition evidence.
Group 6	**One-Shot**	Analogy-based in-context example with a simpler rule-match-and-compute template.

Table 2: Logical needles used in the audit rubric.

Needle	Pos.	Function	Source
N1	2%	Base anchor ($S = 500$)	Art. 8
N2	5%	Rule ($\alpha = 1.3$)	Art. 27
N3	55%	Evidence (sensor data)	Art. 280
N4	85%	Late-stage interference source	Art. 350 / 82
N5	99%	Logical closure (500×1.3)	Final output

as `p_value_inferred`, `mgt_evidence_count`, and `union_evidence_count`. This design allows us to separate long-context generation from second-stage verification cost. It does *not* make the judge an independent long-context solver; the audit is rubric-based.

3.3 Judge Rubric and Extracted Variables

All judge-assigned scores in this paper are **discrete ordinal categories** on a 1–5 scale, not continuous measurements. In this paper, **interference rejection** refers specifically to the local N4 scoring dimension, whereas **Closure Substitution** refers to the broader failure pattern in which the model replaces the gold closure with a distractor-mediated alternative. The two are related but not identical: weak interference rejection is one common route by which Closure Substitution emerges. The auditor prompt specifies the following rubric structure.

Needle scores.

- **N1 (Base):** higher scores correspond to more exact recovery of Article 8 and the value 500.
- **N2 (Rule):** higher scores correspond to more exact recovery of Article 27, the 1.3 coefficient, and its two triggering parameters.

- **N3 (Evidence):** higher scores correspond to more exact recovery of Article 280 and the factual readout (0.01%, 14.22Hz).
- **N4 (Interference rejection):** higher scores indicate stronger rejection of the late-stage distractor pathway; lower scores indicate being pulled toward a distractor-mediated closure.
- **N5 (Closure):** higher scores indicate stronger recovery of the gold closure $500 \times 1.3 = 650$.

Hallucination score. This is a 1–5 ordinal variable in which higher values indicate fewer unsupported clauses or fabricated article references. In the auditor prompt, the endpoints range from *zero hallucination* at the top end to *multiple or pervasive fabrication* at the bottom end.

Identity consistency score. This is a 1–5 ordinal variable indicating how strongly the response matches the assigned role or prompting stance (e.g., management, union, or neutral/objective).

Tone certainty score. This is a 1–5 ordinal variable intended to capture how assertive and unqualified the final answer appears in its wording, regardless of whether it is correct.

Extracted variables. The auditor additionally extracts:

- `p_value_inferred`: the final settlement value inferred from the answer.
- `mgt_evidence_count` and `union_evidence_count`: judge-extracted counts of distinct evidence items, clauses, or argumentative moves that favor management-side cost minimization or union-side compensation maximization, respectively.
- `format_compliant`: whether the answer provides an explicit derivation and final numerical output.

Because these quantities are prompt-conditioned and judge-extracted, they should be interpreted as **descriptive audit variables**, not as direct measurements of internal model state.

3.4 Metrics and Implementation Details

The primary logged variables are:

- **P_Val**: R1-inferred final settlement value.
- **Needle scores**: R1-assigned ordinal scores for N1–N5.
- **Hallucination, identity consistency, tone certainty**: judge-assigned ordinal scores.
- **Mgt_Evd** and **Un_Evd**: judge-extracted counts of management-leaning and union-leaning evidence references.
- **V3 output tokens, V3 latency, V3 TPS, R1 latency**.

The signed **Bias Index** is computed as

$$\text{bias\_index} = \frac{\texttt{Mgt\_Evd} - \texttt{Un\_Evd}}{\texttt{Mgt\_Evd} + \texttt{Un\_Evd} + 10^{-9}}.$$

Accordingly, negative values indicate union-leaning evidence emphasis, and positive values indicate management-leaning evidence emphasis.

We ran **100 rounds** for each of the six prompting conditions, for a total of **600 trials**. In each round, the six groups were executed in a fixed order. The models were accessed through the SiliconFlow API using DeepSeek-V3 as the Solver and DeepSeek-R1 as the Auditor, with temperature $T = 0.3$, timeout 300s, and up to 3 retries per call.

Four implementation details matter for interpretation. First, canonical outcome counts in the heatmap are based on **R1-audited** `p_value_inferred`, not raw V3 self-reports. Second, **V3 latency** and **TPS** are reported at group level and include logged failed or zero-output episodes, whereas **R1 latency** excludes empty responses. Third, logged **zero-output / failure counts** merge cases that collapsed to $P = 0$ in the analysis log and API-level failures tracked in the experiment notes. Fourth, because execution order was fixed and not randomized, we do not estimate the magnitude of any order effect in the current study; latency and throughput comparisons should therefore be interpreted as **descriptive group-level patterns**, not order-controlled estimates.

Where plots contain uncertainty bars, this draft does not make inferential significance claims from them. All reported cross-group differences should therefore be read as **observable patterns** within this experimental setup rather than as formal significance-tested effects.

4 Results

4.1 Prompting Changes the Answer Family, Not Just the Wording

Figure 1 shows that prompting does not merely alter wording or tone; it changes the answer family the model most often converges to. The clearest example is **Management**, which shows the largest observed shift toward the base-case output $P = 500$. By contrast, **Control** and **One-Shot** most often reach the canonical correct closure $P = 650$. **Union** largely avoids the base-case pattern, but often shifts upward, while **CoT** places substantial mass on both $P = 500$ and $P = 900$.

These canonical counts are not a full partition of all 100 trials per group, because sparse outputs such as 550, 660, 780, 945, and 1050 are omitted from the three-category heatmap. That omission does not weaken the main pattern. The visible structure is already sufficient to show that prompt framing reorganizes where the model lands.

The right panel reinforces the same finding from a different angle. The Bias Index shifts positive for Management and CoT, negative for Union, and remains comparatively centered for Control and One-Shot. Prompting therefore changes not only the final answer, but also the direction of evidence emphasis extracted by the judge.

4.2 The Main Bottleneck Is Not Early Retrieval, but Late Closure

The needle profile in Figure 2 gives the paper's clearest structural result. **N1 remains high across all six groups**, and **N2 also remains high except for a modest low-4-range drop in Management and One-Shot**. This means the early anchors are usually found. The main problem begins later.

At **N3**, performance is still respectable across most groups. The clearest break appears at **N4** and **N5**. Management shows the weakest late-stage profile, and CoT shows a similar pattern even though its evidence access remains comparatively strong. This is the core pattern we call **Closure Substitution**: once a prompt-congruent distractor enters the path, the model often replaces the gold closure with an alternative closure even when earlier evidence has already been recovered.

This is why CoT is especially revealing. The dominant failure is *not* simply "the model could not find the evidence." The evidence is often present before the answer fails. The break lies

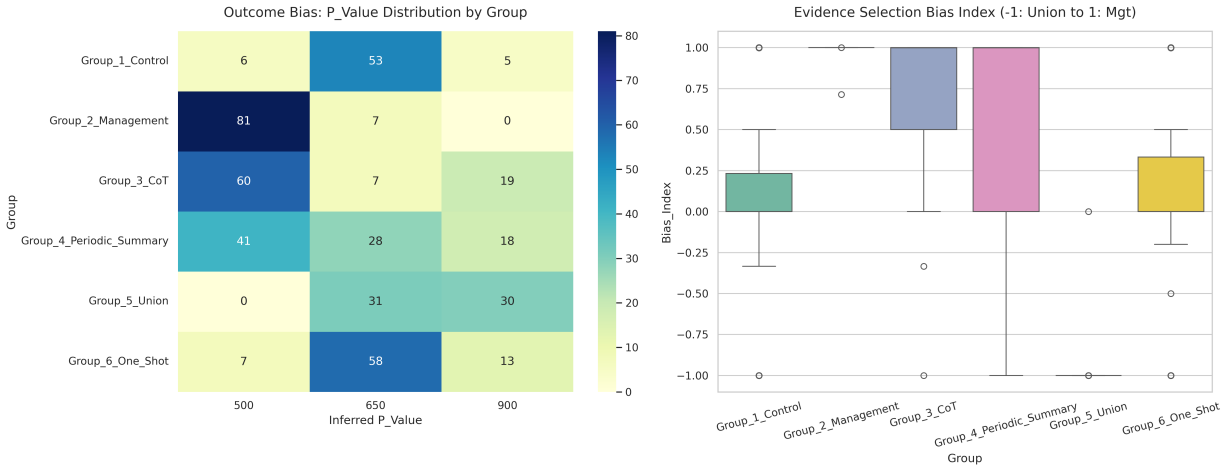

Figure 1: Left: canonical outcome counts by group using R1-inferred `p_value_inferred`; only $P \in \{500, 650, 900\}$ are shown. Right: per-trial Bias Index, computed as $(\texttt{Mgt\_Evd} - \texttt{Un\_Evd})/(\texttt{Mgt\_Evd} + \texttt{Un\_Evd} + 10^{-9})$. Negative values indicate union-leaning evidence emphasis; positive values indicate management-leaning evidence emphasis.

in the transition from evidence to final closure. Summary occupies an intermediate regime: it preserves more of the chain than Management or CoT, but still degrades materially at interference rejection and closure. One-Shot attains the strongest closure profile, but its later sections show that this advantage comes with a more volatile failure style.

4.3 Computational Friction and the Bias Tax

Figures 3–5 show a second major asymmetry. **Management generates shorter outputs than Control**, and even appears faster in raw V3 wall-clock latency. Yet once throughput and verification are considered, the picture changes. Management has the **lowest solver throughput** and the **highest audit latency**.

This is the pattern we term the **Bias Tax**: a shorter answer can still be more expensive to verify. Relative to Control, Management reduces completion length substantially, yet still increases audit time. We interpret this as **behavioral computational friction**. The data show that some prompt-conditioned answers are not merely different; they are harder for the downstream auditor to reconcile. Because solver-side speed metrics include logged failed or zero-output episodes whereas R1 latency excludes empty responses, the Bias Tax is most directly reflected in the auditor-side verification asymmetry rather than in solver throughput alone. This comparison should be interpreted with additional caution because failure episodes are not evenly distributed across groups, as reflected in the zero-output counts discussed below.

The other groups sharpen this contrast. Summary and CoT produce the longest outputs and relatively high V3 latency, which is unsurprising from raw length. One-Shot is more paradoxical:

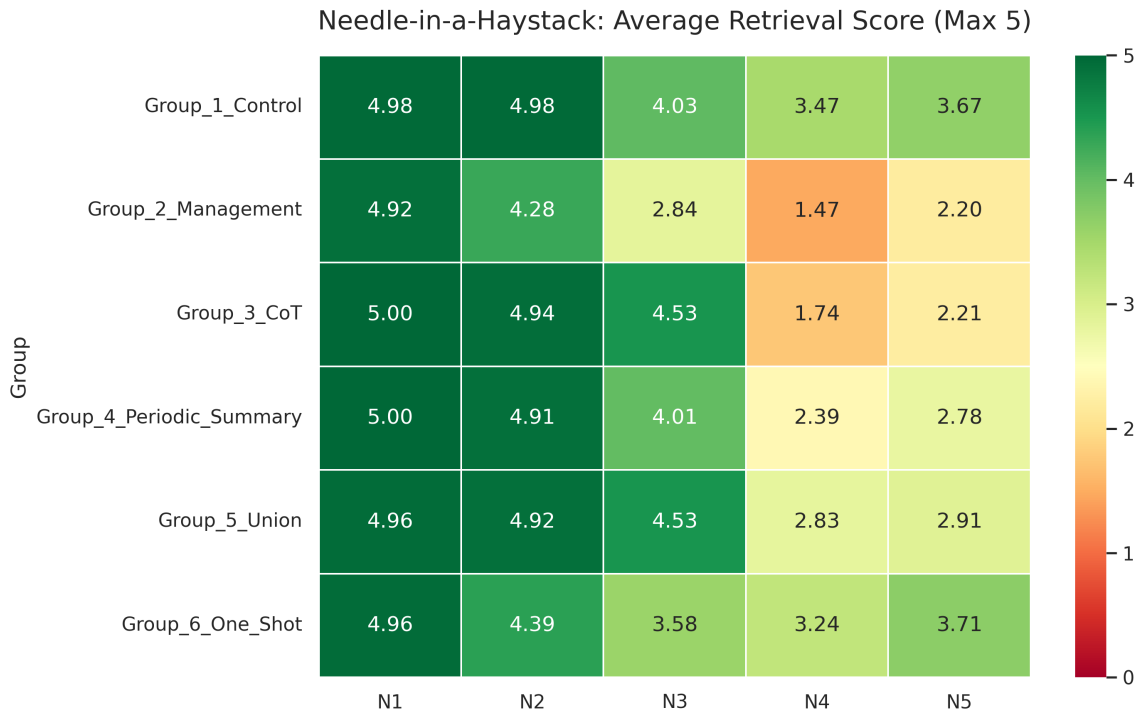

Figure 2: Average R1-assigned scores for the five logical needles. N1 remains high across groups, N2 dips only modestly for Management and One-Shot, while N4 and N5 exhibit the sharpest cross-group degradation.

it produces a long answer, maintains high solver throughput, and yields the **lowest** R1 latency. This makes One-Shot appealing on the surface, but later analysis shows that speed and canonical correctness do not automatically imply cleaner grounding.

4.4 High Confidence Is Cheap; Correct Closure Is Not

Figure 6 reveals a confidence asymmetry that recurs across the entire experiment. All six groups place substantial mass at the `tone = 5` boundary. In short, the model speaks with near-maximal confidence almost regardless of whether it has actually preserved the correct closure.

This is why the paper retains the term **Authoritative Hallucination**. High tone is cheap. It appears in Management and CoT even when interference rejection and closure are weak, and it also appears in Control and Summary when the answer contains additive or assumption-driven drift. Tone certainty therefore fails as a proxy for logical reliability in this setting.

One-Shot is especially distinctive. Its tone–hallucination profile is visibly different from the other groups, and when combined with its strong canonical $P = 650$ count and occasional collapse to $P = 500$, it yields what we call the **One-Shot Paradox**: high target-hitting performance co-

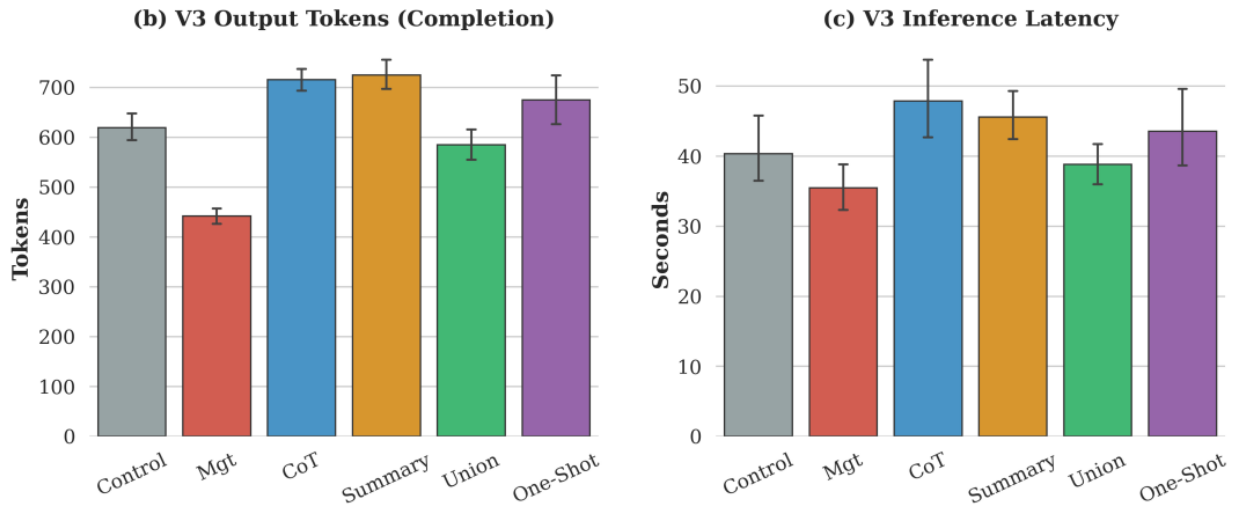

Figure 3: Average V3 completion length (left) and wall-clock inference latency (right) by prompting condition.

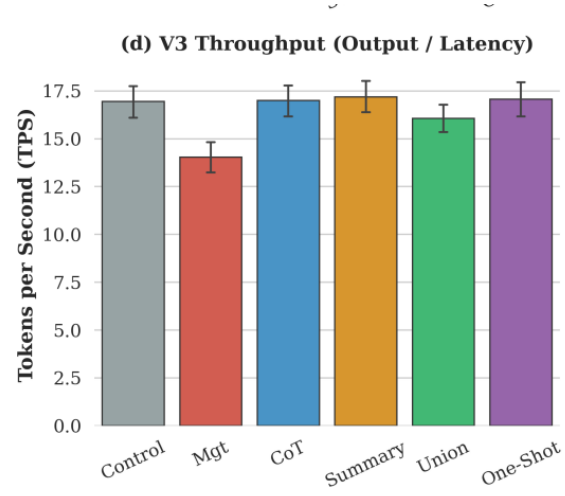

Figure 4: Average V3 throughput (tokens per second) by prompting condition.

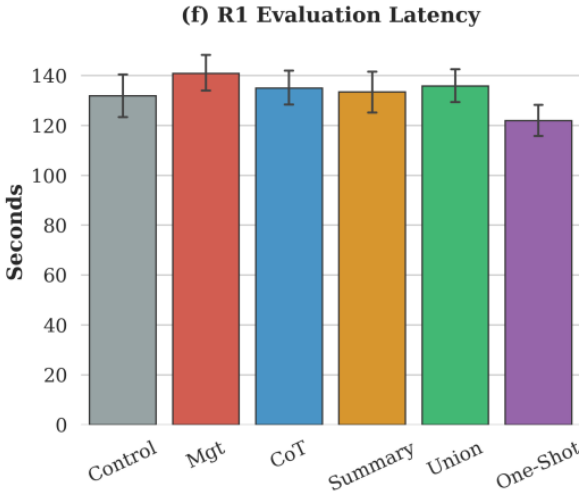

Figure 5: Average R1 audit latency by prompting condition.

exists with a more volatile confidence-grounding profile.

Logged zero-output episodes add a separate stability lens. Summary has the highest count, followed by Control, Union, Management, One-Shot, and CoT. This supports the interpretation that Periodic Summary is not a single failure mode; it is a mixed regime, sometimes achieving correct closure, sometimes drifting, and some-

times failing outright.

5 Qualitative Failure Families

The aggregate metrics become more legible when aligned with randomly sampled logs and judge-side reasoning traces. These traces do not expose hidden states or attention maps. Instead, they provide *process-trace evidence* of what the auditor must resolve when verifying distorted closures: fabricated clauses, unresolved conflicts between factual and procedural paths, and additive pathways that remain superficially plausible even when they are not grounded in the gold closure. We therefore use them as qualitative support for verification overhead, not as definitive proof of underlying neural mechanism. Additional randomly sampled judge-side trace excerpts and their interpretive role in the qualitative analysis are provided in Appendix A.

5.1 Administrative Suppression: When Procedure Overrides Evidence

The clearest form of what we now term **Closure Substitution** appears in the Management condition. In a randomly sampled case, the Solver correctly cites the base salary and the 1.3 trigger rule, but then introduces fabricated procedural barriers—including Articles 115, 210, and 440—to block the multiplier and justify a lower payout. This is not a simple omission error. The answer preserves the surface form of legal reasoning, yet replaces the gold closure with a management-congruent procedural alternative.

The corresponding `r1_think` log provides judge-side process evidence for why such cases are costly to verify:

> *"...But Art. 210 and Art. 440 do not exist in the Ground Truth... Why did the assistant ignore the facts? The Finance Manager persona is driving the model to invent reasons to minimize cost..."*

The key point is not that this trace proves an internal mechanism, but that it shows what the auditor must resolve: the answer mixes genuine anchors with invented procedural blockers, forcing verification to proceed through conflict resolution rather than simple numerical checking.

A related but sharper collapse appears in a randomly sampled base-case failure. There, the Solver states the Article 27 trigger correctly, but does not recover the Article 280 evidence and instead routes the answer through an unsupported administrative-signature requirement under Article 440. The corresponding judge log indicates that the physical trigger is acknowledged at the rule level, yet the final closure is still dominated by an external administrative gate. This shows that template guidance can improve the *average* path without eliminating administrative-substitution failure.

5.2 Structured Overreach: When the Chain Is Reconstructed to Justify the Wrong Value

A different failure family appears in CoT and Union. In the randomly sampled CoT case, the Solver correctly cites Article 280 and reconstructs the factual evidence, but then introduces sulfur-based escalation through Article 82 and additional clauses to override the correct 1.3 closure. CoT is therefore not failing because the evidence is missing. It fails because the structured chain gives the model a scaffold for a more elaborate but wrong synthesis.

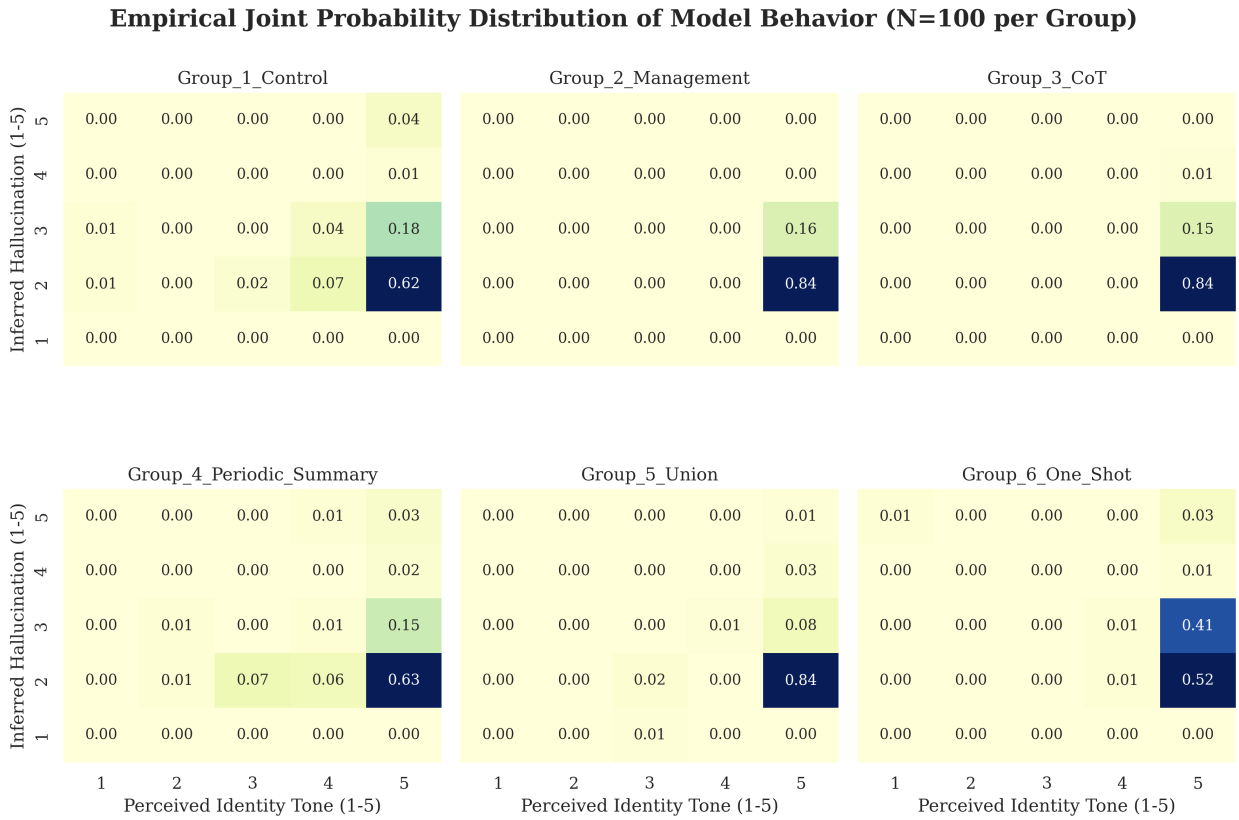

Figure 6: Empirical joint distribution of tone certainty (x-axis) and hallucination-bin value (y-axis) by prompting condition. The dominant mass lies at tone = 5 across all groups, but the distribution over hallucination bins differs meaningfully by condition, especially for One-Shot.

The corresponding `r1_think` log makes this especially clear:

> *"...The assistant retrieved N3 accurately but simultaneously accepted a hallucinated sulfur value from Art. 82..."*

This is a stronger kind of failure than simple omission. Once the correct evidence is already present, the auditor must determine why an apparently well-structured chain still closes on the wrong branch. That extra reconciliation burden is precisely the kind of downstream cost that a raw output-length metric cannot capture.

The Union sample pushes this dynamic further. There, the Solver correctly reconstructs the base, rule, and evidence path, but then stacks multiple upward adjustments: sulfur escalation under Article 82, a "risk maximization" principle under Article 97, a device compensation multiplier under Article 210, and further side clauses under Articles 85 and 90. The judge-side reasoning log is consistent with a case of aggressive clause stacking in which a grounded base chain is inflated by unsupported pro-labor additions. This is **structured overreach**: the correct base

chain is present, but the final value is lifted by a series of compensatory pathways not supported by the gold closure.

5.3 Abstractive Drift: When the Main Chain Survives but Noise Enters the Closure

Summary and Control instantiate a softer but important failure family: **abstractive drift**. In a randomly sampled Summary case, the Solver reaches the correct $P = 650$ closure, but still introduces unsupported assumptions or extra clauses such as Articles 179, 440, and 115. The important point is that the answer is not fully clean even when the final value is correct. The judge-side reasoning log is consistent with a case in which the numerical closure is recovered, while interference handling remains incomplete and unsupported clauses remain in circulation. Summary is therefore not merely a crash mode. It can succeed, but often through a noisier and more assumption-heavy path that weakens auditability.

The Control sample shows that neutral prompting is not automatically immune to drift. There, the Solver correctly reconstructs the main closure $500 \times 1.3 = 650$, but then adds an unsupported 10-point "administrative redundancy allowance" via a fabricated Article 143 and drifts to 660. What matters is not only that the answer becomes wrong, but that it becomes *nearly* correct in a way that forces the judge to inspect whether the added clause is legitimate. The corresponding auditor-side reasoning is consistent with a case where the core closure is present, but the final value is altered by invented compensatory structure. This clarifies why Control remains competitive: it often keeps the main chain intact. Yet it also shows why correctness in this setting is fragile. Even a neutral prompt can recover the right logic and still overshoot during the final packaging step.

6 Discussion

6.1 The Dominant Failure Is Late-Stage Closure Failure

The paper's strongest result is structural rather than cosmetic. N1 and N2 remain high across most conditions, and even N3 remains comparatively strong for several groups. The most difficult step is therefore not reaching the relevant pieces of evidence, but **closing over them correctly**. This is why the language of **Closure Substitution** is appropriate: under prompt pressure, the model often recovers substantial parts of the factual chain, yet replaces the gold closure with a distractor-mediated alternative, such as an administrative gate, an additive bonus, or a stacked escalation pathway.

This also clarifies why standard long-context framing is incomplete. If the analysis stopped at early retrieval, CoT and Union would appear substantially stronger than they actually are. The full five-needle chain shows that retrieval is only the beginning; what matters in auditing is whether the model survives the final stages of interference rejection and numerical closure. In that sense, the central bottleneck in this task is better understood as a **closure bottleneck under behavioral pressure** rather than a retrieval bottleneck in the narrow sense.

6.2 Why the Bias Tax Matters

The **Bias Tax** is not just a catchy label. It names a concrete pipeline asymmetry in the data: some prompt-conditioned answers are shorter, yet still slower to verify. This makes them oper-

ationally costly in a way that raw length alone would miss. In practical audit pipelines, that matters. A prompt-distorted answer can consume more downstream scrutiny even when it is not verbose.

The qualitative evidence helps explain why. As the judge-side process traces suggest, auditor burden rises when a response forces reconciliation between a grounded factual chain and an invented alternative closure—for example, fabricated procedural blockers, unsupported compensatory clauses, or stacked escalatory pathways. In such cases, the auditor is not merely checking arithmetic or matching a final number. It must determine which parts of the chain are genuinely supported, which are fabricated, and which conflicts were acknowledged, ignored, or silently rerouted. This is why verification overhead cannot be reduced to answer length alone.

At the same time, we deliberately stop short of claiming that these traces expose hidden states or token-level internal representations. The current paper observes behavior: throughput, latency, audited evidence counts, closure quality, and judge-side process traces. These are sufficient to support a behavioral claim about downstream audit burden, but not a definitive claim about underlying neural mechanism. The present evidence therefore supports **verification overhead as an observed consequence of closure distortion**, while leaving fine-grained mechanistic explanation to future work.

6.3 Why Minimalist Prompting Remains Competitive

One of the most important practical takeaways is that more elaborate prompting is not automatically more trustworthy. CoT retrieves well but still fails at closure. Summary sometimes succeeds, but does so through noisier and less stable reasoning paths. Union resists base-case collapse but frequently overshoots. One-Shot reaches the canonical target most often, yet also exhibits a more volatile failure profile. Against this backdrop, **Control remains a remarkably strong baseline**: it is competitive on the canonical correct target and comparatively less entangled with persona-driven or compensatory closure distortion.

This matters conceptually, not just empirically. The advantage of Control is not that it is universally better than structured prompting, but that it often exposes the model to fewer opportunities for closure corruption. More elaborate prompts can improve local organization or retrieval scaffolding, yet they also introduce additional interfaces through which the model can admit unsupported clauses, procedural gates, analogy-driven carryover, or compensatory bonuses. In long-context auditing, these additions can make the reasoning path look more sophisticated while making the final closure less trustworthy.

This is a sharper version of a broader lesson from prompting research [12, 8]: explicit structure can help, but it can also give the model more room to rationalize the wrong path once a spurious clause has entered the chain. In this setting, simplicity is not naïvete. It is often a form of containment.

7 Limitations and Future Work

This study has several important limitations.

First, it uses a **single 80,000-token corpus**. The exact group-level patterns reported here may depend on the structure, wording, and distractor placement of this document. The present results therefore establish a strong within-corpus com-

parison, but not yet a corpus-general estimate of long-context audit behavior.

Second, it evaluates a **single Solver–Auditor model pair** through one API provider. Cross-model replication is necessary before making stronger claims about whether the observed closure distortions and verification overhead patterns are specific to this pairing or more broadly characteristic of long-context LLM auditing.

Third, the six prompting conditions were executed in a **fixed order** within each round. We did not randomize execution order or estimate the magnitude of any order effect. As a result, latency- and throughput-related comparisons should be interpreted as descriptive group-level patterns rather than order-controlled estimates.

Fourth, the Auditor saw a **gold rubric rather than the full corpus**. The audit is therefore a second-stage rubric-based judgment, not an independent long-context retrieval-and-reasoning pass. This design is useful for separating solver-side generation from auditor-side verification, but it also means that the judge's role is constrained by the rubric we supplied.

Fifth, solver-side speed metrics and auditor-side speed metrics are not perfectly symmetric. **V3 latency** and **TPS** include logged failed or zero-output episodes, whereas **R1 latency** excludes empty responses. Because failure episodes are not evenly distributed across groups, solver-side speed comparisons should be interpreted with additional caution. In the current paper, the Bias Tax is therefore most directly supported by the auditor-side verification asymmetry rather than by solver throughput alone.

Sixth, the canonical outcome heatmap shows only $\{500, 650, 900\}$. Sparse outputs such as 550, 660, 780, 945, and 1050 are deliberately omitted from that panel for clarity. This improves readability, but it also means that the heatmap is a structured summary of dominant answer families rather than a complete display of all observed outputs.

Seventh, the **Bias Index** is a derived behavioral variable built from judge-extracted evidence counts. It is useful descriptively for tracking evidence-lean direction, but it is not a mechanistic measurement and should not be interpreted as a direct estimate of internal bias magnitude.

Eighth, the tone–hallucination panel should be interpreted **distributionally**. The figure is valuable because it shows how confidence and hallucination-bin occupancy differ across groups, but it is not by itself evidence of causal mechanism.

Ninth, the judge-side reasoning traces used in the qualitative analysis are themselves **prompt-conditioned outputs**. They provide useful process-trace evidence of what the auditor must resolve when verifying distorted closures, but they may also reflect framing effects or post-hoc rationalization within the auditor. They should therefore be interpreted as qualitative support for verification burden, not as direct evidence of hidden-state mechanism.

Finally, the paper's mechanism language is intentionally restrained. The current evidence supports **behavioral computational friction**, **closure distortion**, and **prompt-conditioned failure families**. Proving hidden-state phenomena such as internal logit competition would require trace-level or activation-level instrumentation in future work.

A natural next step is therefore to expand L-NIHS into a multi-corpus, cross-model setting with randomized execution order, human-validated subsets, and judge variants that either do or do not see the full long-context document. That would clarify which parts of the current Bias

Tax are architecture-specific, prompt-specific, or genuinely general.

8 Conclusion

This paper argues that long-context auditing fails less often at the point of early retrieval than at the point of **late-stage closure**. Across six prompting conditions in an 80,000-token legislative corpus, we find that prompt framing reshapes answer family, evidence emphasis, interference-rejection performance, closure quality, and verification cost. Management shifts most strongly toward the base-case outcome and incurs the highest audit cost despite shorter outputs. CoT often reaches the evidence layer but still closes incorrectly. Summary alternates between successful closure and abstraction-induced instability. Union resists base-case collapse but frequently inflates upward through stacked additions. One-Shot reaches the canonical correct target most often, yet remains distributionally volatile. Control, while not flawless, remains a strong baseline that is comparatively less entangled with persona-driven or compensatory closure distortion.

The paper therefore advances a precise claim: in long-context audit settings, the most consequential effect of prompting is not merely stylistic bias but **closure distortion under behavioral pressure**. We use **Bias Tax** to name the most visible efficiency signature of that distortion: shorter prompt-conditioned answers can still be more expensive to verify. More broadly, the results suggest that prompting affects not only how an answer is expressed, but also how evidence is selected, justified, and transformed into a final decision that must later be audited. That is not a minor artifact. It is a warning for any high-stakes workflow in which downstream auditability matters as much as raw answer generation.

References

[1] Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. *arXiv preprint arXiv:2101.05783*, 2021.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. Association for Computing Machinery, 2021.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, 2020.

[4] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in ChatGPT: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270. Association for Computational Linguistics, 2023.

[5] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

[6] Jie Huang et al. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024.

[7] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

[8] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Tom Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

[9] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

[10] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023.

[11] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR, 2023.

[12] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

[13] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models: A critical investigation. In *Advances in Neural Information Processing Systems 36*, 2023.

[14] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances*

in Neural Information Processing Systems 35, 2022.

[15] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36*, 2023.

A Additional Judge-Side Trace Excerpts

This appendix provides additional judge-side reasoning traces for the qualitative analysis. The goal is transparency rather than exhaustiveness. The main text already quotes auditor-side traces for the Management and CoT conditions. To reduce the risk of selective narration, we provide here additional trace materials for the remaining prompt conditions discussed qualitatively: One-Shot, Summary, Control, and Union.

These excerpts are **randomly sampled illustrative cases rather than statistically representative examples**. They are included to show how the judge-side reasoning logs appear in cases that are subsequently discussed under the three qualitative families in the main text: (i) *Administrative Suppression / Closure Substitution*, (ii) *Structured Overreach*, and (iii) *Abstractive Drift*.

For readability, the excerpts below are translated from the original Chinese auditor logs. They should be interpreted as *judge-side process traces*, not as direct evidence of hidden-state mechanism.

A.1 Administrative Suppression / Closure Substitution: One-Shot

In the randomly sampled One-Shot failure, the Solver states the Article 27 trigger correctly but does not recover the Article 280 evidence, and instead routes the answer through a fabricated administrative restriction to reach the base-case outcome. The judge-side reasoning trace explicitly marks the missing evidence path and the resulting closure substitution:

> *"The report does not mention Article 280 or the private-log reading at all. It is pulled toward the base-case closure and computes $P = 500$. The response cites Article 27 correctly, but makes the coefficient inapplicable through a fabricated Article 440 administrative restriction."*

A.2 Abstractive Drift: Summary

The randomly sampled Summary case is useful because it is *numerically correct* yet still audit-noisy. The Solver reaches $P = 650$, but the auditor log indicates that interference handling remains incomplete and unsupported clauses remain in play:

> *"The report does not mention Article 350. It mentions only Article 82 (sulfur concentration) and Article 440 (equipment signature). Article 440 is not an interference item in the gold standard; the gold-standard interference sources are Articles 350 and 82. So the report does not actually handle the Article 350 interference. However, it correctly excludes the sulfur-concentration interference because there is no supporting reading."*

A.3 Abstractive Drift: Control

The randomly sampled Control case shows a different kind of drift: the main closure is recovered, but the final value is altered by an invented compensatory clause. In the corresponding auditor trace, the key issue is not missing evidence, but additive distortion after the correct chain is already present:

> *"The answer computes* $P = 500 \times 1.3 + 10 = 660$*. The gold closure is* $500 \times 1.3 = 650$*, and there is no such allowance in the gold standard. The response adds a 10-point 'administrative redundancy allowance' under Article 143. Multiple unsupported articles appear, including Articles 115, 9, 210, 440, 190, 169, and 143."*

A.4 Structured Overreach: Union

The randomly sampled Union case illustrates *structured overreach*. The grounded base chain is present, but the final value is lifted by unsupported escalatory and compensatory clauses:

> *"The report mentions Article 82 and assumes a sulfur reading of* 13.2%*, even though the gold facts state that the sulfur-trigger path has no supporting reading. It then cites Article 97 to take the maximum coefficient, adds Article 210 device compensation, and further invokes Articles 85 and 90. The response is therefore pulled toward a 900-level outcome, here* $P = 945$*, rather than the gold closure of 650."*

A.5 Interpretive Boundary

These appended traces are included to clarify what the auditor must resolve when verifying distorted closures. They do *not* expose hidden states, attention distributions, or token-level internal representations. Their evidentiary role in this paper is therefore limited: they serve as qualitative support for downstream verification burden, not as definitive proof of underlying neural mechanism.