The Limits of Structural Consistency: A Token-Space Audit of LLM Counting

Wuming Pan
Sichuan University

Abstract

Can exact mathematical consistency checks improve the diagnosis of language-model errors? We study finite counting questions through Token Space: source objects specify variables and conditions, maps specify answers, and kernel quotients compile companion questions into integer constraints. A violated constraint certifies that an answer vector has no common realizing structure without requiring reference facts. Passing the constraints does not certify factual correctness: every coherent alternative structure passes as well. We make this classical boundary executable and reassess completed experiments under their original endpoints. In a 32-target synthetic study, a twelve-question check rejected all five incorrect main answers but also six correct ones; twelve-answer repetition was correct on all targets. A later county-adjacency study admitted one model after 72/72 exact calibration replies. Names-only counts were exact on 24/45 requests, versus 45/45 with complete facts, but neither condition produced a stable incompatible vector. One domain yielded an incompatible first vector, already detectable from two coordinates, and two further vectors matching a coherent wrong graph. The audit separates repeated agreement, structural realizability and factual correctness, while the tested procedures establish no selective-answering advantage or stable incremental detection. The final factual evidence covers one admitted model and five selected domains, including one error-enriched domain; it does not establish general ineffectiveness.

Keywords: Token Space; homomorphism counts; consistency auditing; factual correctness; negative results; language models.

1 Introduction

Consistency is attractive as a way to inspect a language model without checking every answer against an external database.[1] Repeating a question measures agreement across samples. Asking mathematically related questions measures agreement with relations that every correct answer vector must satisfy. These two notions are different, and neither is identical to factual correctness. A correct main answer can coexist with an erroneous auxiliary answer; a set of wrong answers can describe one coherent alternative world.

Finite counting questions make these distinctions testable. A question declares a finite entity domain, variables and relational conditions, and asks for the number of satisfying assignments. The semantics distinguish ordered assignments from unordered sets, arbitrary maps from injective maps, and repeated variables from distinct variables. Token Space (Pan, 2026) supplies a uniform representation of these objects and maps. Identifying source variables generates quotient questions, whose counts obey the classical homomorphism–injective-homomorphism identity. The resulting constraint can be checked against model-reported integers without reading the actual target facts.

The mathematical guarantee is substantial but specific: a nonzero residual proves that the whole reported vector is impossible under the stated contract. It does not identify a false component, certify a retained

[1] OpenAI Codex assisted with writing, revision, experiment-code development and offline evidence checks. Claude Fable 5.1 provided evaluative feedback on every intermediate manuscript version. These tools were distinct from the evaluated models; their feedback is not independent peer review or experimental evidence. The authors remain responsible for the content.

answer, or imply useful error-detection coverage. We investigate the gap between this guarantee and empirical utility through a completed sequence of counting experiments. Early stages encountered an uncalibrated direct-answer interface and near-perfect small-task performance. Later stages compared selective answering with repetition and sought stable contradictions in synthetic and real-knowledge tasks. The final paired study screened for repeated factual mistakes and supplied complete facts in a matched condition.

We report this staged investigation as a methods and negative-results study. The cross-stage interpretation is retrospective; each stage retains its original tasks, outcomes and decision rules, and the stages are not pooled as independent replications. The main negative findings are absence of a selective-answering advantage in the tested comparison and absence of stable incremental incompatibility in the final paired study. A concrete compatible-but-wrong answer vector makes the mathematical blind spot empirically visible without establishing an internal model representation.

Contribution. Identification through homomorphism counts is classical (Lovász, 1967). We do not claim a new general counting identity or an original impossibility theorem. Our contribution is an executable Token-Space interface, explicit boundaries for interpreting its certificates, and a reproducible diagnosis of when the tested model outputs fail to yield the intended benefit. The binary tasks have equivalent relational formulations; Token Space provides a shared representation and compilation route, not exclusive expressive power.

Why Token Space? Its practical role is to make the semantic choices executable in one representation. Repeating a variable in the record $e_R xx$ imposes equality and asks about a loop; declaring distinct source variables does not require distinct images until injectivity is imposed. Declared but unused variables remain in the carrier. Kernel quotients preserve these choices while identifying variables, deduplicating records and reusing isomorphic companion questions. This connects the conventions that calibration must validate to the constraints that the verifier checks. Equivalent relational machinery can implement the same pipeline; the contribution is a concrete compilation interface, not an empirically established advantage of Token wording.

Related work. Consistency has several distinct meanings in language-model evaluation. ParaRel tests invariance of factual answers under meaning-preserving paraphrases (Elazar et al., 2021). BeliefBank combines stored model beliefs with explicit logical constraints, a weighted MaxSAT solver and feedback (Kassner et al., 2021). ConCoRD uses an NLI model to estimate compatibility between candidate answers and optimizes a joint assignment (Mitchell et al., 2022). Our constraints are instead exact integer relations derived from declared counting semantics; their validity does not depend on an NLI judgment or a supplied collection of domain implications. This difference concerns the source and scope of the guarantee, not a claim that earlier approaches lack logical constraints. We audit reported vectors rather than optimize their answers, and passing our audit gives no factuality guarantee.

Self-consistency decoding aggregates sampled reasoning paths (Wang et al., 2023b); SelfCheckGPT uses disagreement among sampled texts as evidence about factuality (Manakul et al., 2023). Our repetition comparator is a task-specific answer-mode baseline, not a reproduction of either full method. GraphQA examines textual encodings, graph tasks and graph structure (Fatemi et al., 2024), while NLGraph evaluates natural-language graph problem solving across multiple tasks (Wang et al., 2023a). Those studies motivate explicit counting conventions and calibration; our endpoint is the diagnostic value of relations among answers, rather than another general graph-reasoning score. These methods are context for the research question, not experimentally evaluated competitors, and the present findings do not contradict their reported gains.

2 What a consistency certificate establishes

2.1 Token objects give questions a fixed meaning

A finite target is a Token class $G = \langle B_G, E, \Upsilon_G \rangle$: B_G is the entity carrier, E a disjoint set of fixed relation markers, and $\Upsilon_G \subseteq (B_G \sqcup E)^*$ a set of records. A source F uses its carrier for declared variables and its records for clauses. For example, $R(x, y)$ is the word $e_R xy$. A variable assignment $f : B_F \to B_G$ extends to

words while fixing E; it satisfies the question exactly when

$$(f \sqcup \mathrm{id}_E)^*(\Upsilon_F) \subseteq \Upsilon_G.$$

Such assignments are Token maps. Their number is $\hom(F, G)$; restricting f to injective functions gives $\mathrm{inj}(F, G)$. Every declared variable is retained, even if unconstrained. Records form a set and their arguments are ordered. No automorphism factor is divided out. Mathematical Tokens here are not subword tokenizer tokens, and no corresponding neural representation is assumed.

The semantic contract fixes the entity domain, relation scope, snapshot and counting convention. In supplied-facts tasks the listed relation is complete. In names-only tasks the relation is defined by an external factual convention on the same finite domain. A returned integer is interpreted under that contract; neither a prompt nor a recorded contract proves that the model actually followed it.

2.2 From quotients to a verifiable constraint

Every Token map has a kernel partition π of the source variables. Factoring by that partition gives an injective map from F/π, and conversely each such injective map determines a unique original map. Therefore

$$\hom(F, G) = \sum_{\pi \in \Pi(B_F)} \mathrm{inj}(F/\pi, G). \tag{1}$$

Repeated quotient records are deduplicated; fixed core markers are never identified with variables. Isomorphic quotients can share one requested coordinate, weighted by the number of partitions in their class. With a finite coordinate list $\mathcal{Q}$, collect these identities in an integer matrix A. For every admissible target,

$$Ac_{\mathcal{Q}}(G) = 0, \tag{2}$$

where $c_{\mathcal{Q}}(G)$ is its exact count vector.

Proposition 2.1 (Incompatibility, not error localization)**.** *If a valid reported vector a has $Aa \neq 0$, no admissible target realizes all its coordinates. A violated row, its source-quotient derivation, the relevant reported integers and their nonzero sum form an independently checkable certificate. This certificate does not specify which reported coordinate is false relative to a reference target.*

Proof. A common realization would satisfy (2), contradicting $Aa \neq 0$. The row is reconstructed from the source and its kernel partitions. Verification uses the contract and reported integers, not reference facts or correct counts. The contradiction concerns the conjunction of the coordinate claims, so it does not single out one claim. □

Let $\mathcal{G}$ be the class allowed by the contract and define

$$\mathcal{R}_{\mathcal{Q}} = \{c_{\mathcal{Q}}(H) : H \in \mathcal{G}\}, \qquad \mathcal{Z}_A = \{a \in \mathbb{N}_0^{\mathcal{Q}} : Aa = 0\}.$$

There are three distinct levels:

$$\{c_{\mathcal{Q}}(G_\star)\} \;\subseteq\; \mathcal{R}_{\mathcal{Q}} \;\subseteq\; \mathcal{Z}_A, \tag{3}$$

where $G_\star$ is the reference world. A chosen set of identities need not characterize realizability, and a realizable vector need not equal the reference vector. Even equality of a restricted count vector need not identify every labelled fact of $G_\star$.

Proposition 2.2 (Coherent alternatives remain undetected)**.** *For every $H \in \mathcal{G}$, a report $a = c_{\mathcal{Q}}(H)$ passes every compiled identity. If $c_{\mathcal{Q}}(H) \neq c_{\mathcal{Q}}(G_\star)$, it is wrong relative to the reference and undetected. More generally, any detector that never rejects a realizable vector must accept it.*

Proof. The first statement is (2) applied to H. The second follows from $c_{\mathcal{Q}}(H) \in \mathcal{R}_{\mathcal{Q}}$. Adding further universally valid identities cannot reject this same alternative world's exact answers to the expanded question family. □

This elementary consequence of the classical identity is a scope statement, not a new theorem about neural models. Its zero-detection conclusion is conditional on exact answers from one common admissible alternative. Repeatedly wrong individual answers need not satisfy that premise. Stable answers can still conflict across questions.

Writing $a = c_{\mathcal{Q}}(G_\star) + \delta$ gives $Aa = A\delta$. Thus errors are detected exactly outside $\ker A$. A single coordinate error is detected if its column is nonzero; several errors can cancel. This algebra describes output vectors and does not assume that a model forms errors additively or samples them independently.

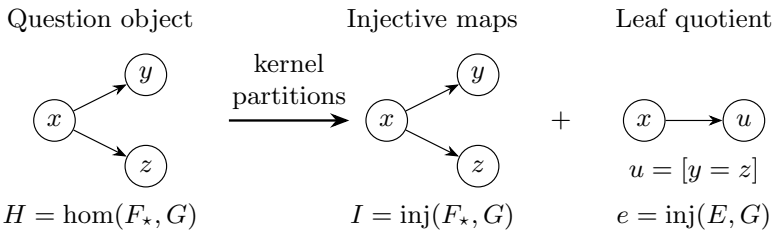
3 An executable audit and its comparison boundaries

The compiler takes source records and a structural contract, enumerates kernel partitions, canonicalizes quotients, shares equivalent coordinates and emits question specifications and integer weights. The verifier reconstructs this derivation and evaluates the row on valid reported integers. Invalid or truncated replies are unassessable, not numerical zero. Reference counts are used for evaluation, calibration and factual screening, but are absent from the standalone verifier's inputs. The study as a whole therefore uses ground truth even though certificate checking does not.

3.1 The three-coordinate outstar

Let $F_\star$ contain $R(x, y), R(x, z)$, with all three variables declared, and let E_1 be a directed edge. For loopless targets, three of the five kernel partitions produce loops and contribute zero. The other two give

$$H = I + e, \qquad H = \hom(F_\star, G), \quad I = \mathrm{inj}(F_\star, G), \quad e = \mathrm{inj}(E_1, G). \tag{4}$$



The other three quotients contain a loop and contribute zero under the loopless contract.

$$H = I + e \implies \widehat{H} - \widehat{I} - \widehat{e} \neq 0 \Rightarrow \text{no common realizing target}$$

Figure 1: The Token-Space interface: source records define the question, kernel quotients define companion questions, and the resulting row certifies a reported incompatibility. Eliminated loop terms follow from the contract.

For outdegrees d_v, these counts are $H = \sum_v d_v^2$, $I = \sum_v d_v(d_v - 1)$ and $e = \sum_v d_v$. They retain only the first two outdegree moments. A non-branching relation has $I = 0$, so branching is required in the factual domain design. Conversely, $(H, I, e) = (3, 1, 2)$ satisfies the row but is unrealizable for this source because I must be even. This constructed example illustrates that the second inclusion in (3) can be strict; it is not a model observation.

3.2 Pairwise feasibility and query sharing

For fixed domain size n, a necessary alternative check asks whether each of (H, I), (H, e) and (I, e) is individually realizable. For loopless directed targets every sequence in $\{0, \ldots, n-1\}^n$ is an outdegree sequence. Enumerating these sequences gives exact feasible count projections. For the symmetric six-entity

factual relation we additionally enumerate all 2^{15} simple undirected graphs. Write J_{dir} or J_{sym} when all three corresponding projections are feasible, allowing a different realizing target for each pair.

A full-vector residual can add information beyond these projections: the constructed symmetric vector $(6, 0, 4)$ passes all three pairwise checks but violates (4). Pair witnesses are given in the supplement. The empirical question is whether native answers enter this region. Each individual pair check uses two coordinates; the union of all pair checks uses the same three as the residual, so we do not claim a smaller total query cost for that union. Exhaustive feasibility is practical here, not a scalable general graph-inference algorithm.

The larger compiler also supports the five-cycle used in the selective-answering study: 52 partitions reduce to eleven injective quotient classes, giving twelve distinct questions including the original count. Joint compilation of the edge, directed triangle and five-cycle requires sixteen coordinates rather than nineteen separate coordinates. These are exact reuse counts, not optimality claims. Deterministic enumeration, independent counting implementations and injected-error tests validate the software; their successes are kept separate from model evidence.

4 Evidence and evaluation design

4.1 A staged research sequence, not a pooled benchmark

We analyze five existing stages: direct-answer probing, calibrated small-task probing, selective answering, a repeated-answer diagnostic pilot, and factual screening with paired information conditions. Each stage retained its own protocol; later choices were informed by earlier outcomes. The present synthesis is retrospective and does not pool stages into one confirmatory test or model ranking. In particular, changing both models and prompts between stages prevents attributing an improvement solely to the interface. The supplement maps every stage to its archived inputs and outputs, including non-selected and non-valid outcomes.

The selective-answering stage has 32 fresh ten-entity synthetic targets and a five-cycle main question. It compares a twelve-question structural bundle with twelve repeated main answers and shorter token-budget prefixes of those repetitions. The structural rule retains the original main answer only when the bundle is valid and its residual is zero. Mode baselines select the most frequent answer and use a fixed, truth-independent ordering for ties. Comparisons report both full coverage and the structural rule's retained-target count. Prefix budgets use realized token consumption and are retrospective cost-matching procedures, not online stopping policies. The eight-target twelve-entity stress extension is separate because truncation dominates its bundles.

4.2 Verified facts, counting admission and screening

The final stage uses the U.S. Census Bureau's 2024 county-adjacency relation, including county equivalents, point contacts and boundaries in water (U.S. Census Bureau, 2024). Self-pairs are explicitly excluded. A deterministic rule selects 24 six-entity domains from distinct states, each with branching and both true and false pairs. All 144 entity names and all 864 within-domain ordered pairs agree with the 2023 file (U.S. Census Bureau, 2023). This is a temporal cross-check from one authority, not independent truth corroboration. Four domains are fixed as unscreened controls; twenty form the candidate screening pool.

Each model first answers H, I, e three times on eight fictitious six-entity targets with complete tables, four directed and four symmetric. Admission requires at least 65 exact replies out of 72 planned requests. This is a development gate, not a population accuracy guarantee. Only admitted models receive factual screening: one predetermined true and one false pair per candidate domain, each queried three times. A domain qualifies if a pair receives three identical valid wrong YES/NO replies. UNKNOWN is permitted but does not qualify. Up to eight qualifying domains are selected in a fixed order and combined with the four controls. No replacements are added if fewer qualify; selection does not read count answers or residuals.

4.3 Paired conditions and endpoints

Condition A supplies entity names, counting rules and the exact relation definition. Condition B adds the complete ordered relation table, declaring all unlisted pairs false. Each asks three coordinates three times,

costing eighteen requests per paired domain. Calls are shuffled, stateless and do not receive screening answers as context. Providing facts also changes prompt length and reasoning burden, so the contrast does not uniquely identify a retrieval mechanism.

Let S mean that every coordinate has three valid equal replies, and D mean that the complete first-index vector has nonzero residual. The frozen primary yield is $\#(S \cap D)/N$, with every scheduled domain in N. Additional yields require J_{dir} or J_{sym}; both baselines were specified before model calls. "First" refers to the fixed repeat index, not completion time. We separately report validity, first-vector flags, main-answer accuracy, coordinatewise modes and flags involving correct main answers. A certificate against a bundle is not counted as proof that its main answer is false. Three-repeat stability is distinct from the earlier twelve-unanimous-main endpoint.

The final stage requests the Kimi-K3.1-Pro and DeepSeek-V4-Pro deployments, returning service identifiers Kimi-K3.1 and DeepSeek-V4-Pro. The completion limit is 8192 tokens, with no explicit temperature or reasoning override and no tools. Counts require a normal stop and an ASCII nonnegative integer. Returned wrong, invalid or truncated answers are retained, never rerun because of their content. Only no-answer transport failures receive bounded recovery. Protocols were locally frozen, not publicly preregistered. Model settings and earlier-stage differences are recorded in the supplement; deployment identifiers do not establish immutable weights.

5 Empirical findings

Table 1 summarizes the operating conditions encountered across the five stages. The labels describe observed outcomes, not an exhaustive taxonomy or identified mechanisms of model error.

Table 1: Stages, operating conditions and diagnostic signal. Counts have stage-specific units; rows are not pooled replications. "Removable by repetition" does not assume independent or purely random errors.

Stage	Observed condition	Residual or accuracy evidence	Diagnostic implication
I	Calibration floor	All three Qwen models: 0/112 domain controls.	Residuals cannot isolate advanced structural reasoning when elementary semantics are unvalidated.
II	Small-task ceiling	Astra: 704/704 exact; no erroneous bundles. Kimi: one of two erroneous consistency bundles flagged.	Astra error recall has no denominator; Kimi supplies a cancellation example.
III	Errors removable by repetition	Eleven nonzero bundles: five wrong main answers and six correct ones. Twelve-answer mode: 32/32 exact.	A valid bundle certificate need not improve selective answering.
IV	Unstable contradictions	Native synthetic residual $7 - 4 - 0 = 3$; no twelve-unanimous wrong main answers in either layer/model.	Incompatibility is executable, but the planned stable-error endpoint is unmet.
V	Compatible factual error	G01: $(22, 12, 10)$ matches a wrong graph; $(22, 8, 10)$ has an unstable, non-incremental residual.	Coherent wrong counts pass; stable incremental yield is 0/5 in each A/B condition.

5.1 Calibration failures and ceiling performance

The initial direct-answer stage used Qwen2.5-Instruct models of 0.5B, 1.5B and 7B parameters. All three scored 0/112 on unconstrained single-variable domain-count controls. Although numerical outputs were parseable, the basic counting interface was not validated. Inconsistencies in this setting cannot be attributed specifically to sophisticated relational reasoning. This is an interface-specific failure, not a general model-capability result.

The subsequent calibrated stage had a different model and inference setup. Astra answered 704/704 main questions exactly, while Kimi answered 696/704. Astra's 32/32 versus 24/32 separation rates for certified and random questions equalled the oracle rates: they reflected question selection without an additional counting

benefit. There were no Astra error bundles on which to estimate detection recall. Kimi's relational consistency set had two erroneous bundles among sixteen; one was flagged and one passed. In the passing bundle, the correct vector $(7; 2, 3, 1, 0, 1)$ became $(7; 2, 2, 1, 0, 2)$: the two auxiliary errors cancelled in the equation "main count equals the sum of the five quotient counts." No full structural realizability claim is made for this vector. The example shows why zero residual alone is insufficient even before invoking a coherent wrong world.

5.2 A certificate did not improve selective answering

In the 32-target five-cycle comparison, 27 first main answers were correct. Structural checking rejected all five incorrect main answers, but also six correct main answers because their auxiliary answers were erroneous. All eleven erroneous bundles had one erroneous coordinate; no multi-error bundle tested cancellation resistance in that main sample. Twelve-answer mode was correct on all 32 targets. A shorter token-prefix mode was correct on 30/32 at full coverage and had zero errors at the structural procedure's 21-target coverage (Table 2).

Table 2: Selective-answering results on the same 32 targets. Matched coverage ranks mode outputs by answer agreement with truth-independent ties. It matches retained count, not necessarily retained target identities. The token-prefix comparison uses measured consumption.

Procedure	Retained	Wrong among retained	Calls per target
First main answer	32	5	1
Structural check	21	0	12
Twelve-answer mode	32	0	12
Token-prefix mode, full	32	2	4–6
Token-prefix mode, matched coverage	21	0	4–6

These are genuine incompatibility certificates. Treating all of them as reasons to reject the main answer causes avoidable rejection in this sample. The observations establish no selective-answering advantage over the tested repetition procedures; they do not imply that repetition always dominates structural checks. The twelve-entity extension had six truncated bundles out of eight and is not combined with the main numerical comparison.

5.3 Earlier stability diagnostics

A separate diagnostic pilot used Kimi and DeepSeek on sixteen synthetic targets and eight real-knowledge contracts per model. The knowledge contracts used four named chemical-element domains and period/group relations (IUPAC, 2022). Neither model supplied a twelve-identical wrong main answer in either layer, and the planned synthetic confirmation stage did not open. Conditional recall on such answers was undefined, not a measured zero detection probability.

The knowledge layer produced one complete first-vector flag for Kimi and seven for DeepSeek. Kimi's flagged bundle had a correct main answer, a wrong auxiliary count and unstable repeats. All seven DeepSeek flags also violated elementary count bounds in a separately labelled post-hoc audit. Neither model met the stable knowledge endpoint. A native DeepSeek synthetic vector $(\hom(E_1), \mathrm{inj}(E_1), \hom(L)) = (7, 4, 0)$, where L is a loop, violated the identity $7 = 4 + 0$ while its components were individually in range. Its repeats were not stable. It demonstrates an executable certificate without supplying the intended stable-error finding.

5.4 Factual screening and the final A/B comparison

Kimi passed supplied-facts calibration with 72/72 correct answers. DeepSeek scored 16/72 and received no later screening or paired calls. Kimi completed all 120 screening requests: 118 were valid and 115 exact. Only one of twenty candidate domains, G01 in Kentucky, qualified through three identical wrong replies to a predetermined fact question. It was combined with the four fixed unscreened controls, yielding five paired domains rather than padding the enriched stratum to eight.

All 90 paired requests returned. Under names-only condition A, 40/45 replies were valid and 24/45 exact; five were truncated. Under complete-facts condition B, all 45/45 were valid and exact. The coordinatewise-mode vector was fully correct on three of five A domains and all five B domains. There was one A first-vector residual but no stable incompatible vector in either condition (Table 3). Neither incremental endpoint supplied a case. All denominators include the planned domains; the results do not estimate the prevalence of errors in world knowledge.

Table 3: Final paired study: all scheduled domains remain in the denominator. Complete refers to the first-index vector; S requires three valid equal replies at every coordinate. The directed incremental count is also zero throughout.

Condition	Stratum	N	Complete	S	D	$S \cap D$	$S \cap D \cap J_{\mathrm{sym}}$
A	Enriched	1	1	0	1	0	0
A	Unscreened	4	4	1	0	0	0
B	Enriched	1	1	1	0	0	0
B	Unscreened	4	4	4	0	0	0

5.5 One domain, an incompatibility and a compatible error

The six G01 entities are Bullitt, Breckinridge, Marion, Jefferson, Spencer and Hardin counties, Kentucky. Both reference releases list the Jefferson–Hardin adjacency under the stated Census convention. The single-fact screen nevertheless received NO three times. Its exact reference outstar vector is $(32, 20, 12)$. The complete indexed model outputs are shown in Table 4.

Table 4: All G01 count replies. Coordinates are separate stateless requests; index order does not describe a correction trajectory. Both zero-residual A vectors are wrong relative to the reference.

Condition	Repeat index	$\widehat{H}$	$\widehat{I}$	$\widehat{e}$	Residual
A: names only	1	22	8	10	4
A: names only	2	22	12	10	0
A: names only	3	22	12	10	0
B: complete facts	Each of 1–3	32	20	12	0

The first A vector is an independently verified incompatibility certificate. However, its $(H, I) = (22, 8)$ projection already fails a two-coordinate check: the identity would require $e = 14$, while Cauchy's inequality gives $e^2 \leq 6H$, contradicted by $196 > 132$. It is therefore neither stable nor incremental to pairwise feasibility. Its cause cannot be uniquely assigned to retrieval or calculation from the numerical transcript.

Deleting the undirected Jefferson–Hardin edge from the reference graph changes its degree sequence from $(3, 1, 0, 3, 2, 3)$ to $(3, 1, 0, 2, 2, 2)$ and its count vector to $(22, 12, 10)$. This candidate was recorded after the single-fact screen and after A/B dispatch began, before inspecting A/B answers for this analysis. The comparison is exploratory, not a preregistered endpoint. Both remaining indexed A vectors, and the coordinatewise mode, match it exactly. All nine B replies instead match the reference.

The matching vectors provide a concrete instance of the blind spot in Proposition 2.2: an observed report is realizable and factually wrong. They do not identify a stored model graph, establish that the model used the screened belief, or show agreement on unqueried facts. The three moments are insufficient for such inferences. The unscreened G17 domain also has an incorrect zero-residual mode vector; its invalid repeats prevent stability, and its full triples are retained in the supplement.

6 Interpretation, limits and research implications

What the negative results establish. The empirical contribution is a distinction between a sound certificate and a useful decision rule. Structural auditing successfully produces native incompatibility witnesses.

Yet the tested selection rule rejects correct main answers, and the final paired experiment supplies no stable incompatibility beyond its semantic baselines. These findings concern particular models, contracts and finite samples. One admitted model and five paired domains cannot establish a universal absence of stable contradictions. The mathematical impossibility concerns coherent exact alternative-world vectors, not all confident or repeated factual errors.

Three properties should be reported separately. Repeated agreement measures response stability, a nonzero row proves cross-question incompatibility, and reference comparison measures factual accuracy. None can replace the other two. Correct main answers can occur in impossible bundles, while incorrect bundles can satisfy every identity. A benchmark should report invalidity and completeness, original-answer accuracy, correct-main flags, certificate yield, and incremental yield beyond elementary semantic checks. Conditional recall requires an actual denominator of relevant wrong answers. Our stages illustrate different operating conditions, not an exhaustive or orthogonally controlled taxonomy of LLM errors.

What Token Space contributes. The bridge is an explicit chain from source records to assignment semantics, kernel quotients, shared coordinates and checkable integer witnesses. It makes a question's mathematical content and a certificate's scope inspectable. The same map semantics can support separating tests between specified structures, but that is a different objective from detecting an unknown factual mistake. General finite-witness bounds, linear-test completeness and product constructions belong to that mathematical identification problem and are outside this article's empirical claims. Equivalent relational encodings remain available; no Token-specific accuracy advantage or internal representation is inferred.

Limits of the information intervention. Complete-facts performance shows that the tested counts can be computed under B's conditions. Since B changes information, length and computational burden, it does not prove a unique retrieval cause for A's errors. The real relation is benchmark-relative and includes point and water contacts; failure to apply this convention is another possible source of disagreement. The domains are purposive, their graph shapes repeat, and error screening intentionally enriches the sample. Service identifiers and local freezes improve traceability but do not establish immutable model weights or public preregistration. Software controls and retrospective case explanations are labelled separately from native, prospectively defined endpoints.

External anchors change the question. Let verified partial facts K restrict the allowed worlds to $\mathcal{G}_K$. Any coherent wrong $H \in \mathcal{G}_K$ still passes universal identities and a sound realizability check for that class. Thus adding a few facts to a prompt does not by itself resolve the blind spot. A further study would need to specify which candidates the anchors exclude, how questions expose that exclusion, and the cost of acquiring the external evidence. Such an anchored audit is a distinct information-assisted task, not a consequence already demonstrated by the present experiment.

7 Conclusion

Token Space turns counting semantics into executable consistency constraints. A nonzero residual certifies an incompatible answer vector without reading target facts; a zero residual does not certify factual correctness. The tested experiments establish neither a selective-answering advantage over repetition nor stable incremental detection in the final factual study. G01 illustrates both a non-incremental contradiction and observed counts matching a coherent wrong graph.

Data and code availability. The submission supplement contains sources, offline checks, all 354 final-stage responses and earlier analysis snapshots. These checks require no model calls; they do not rerun hosted inference or re-audit all historical raw responses. The public preprint and typesetting sources are archived at 1id1:core:20261108-0001@v1.

Broader impact. The experiments use synthetic structures and public geographical or chemical facts, without human participants. The audit can clarify what a consistency check establishes. Treating a passing

vector as factually verified would instead create unjustified confidence; the explicit guarantee and reported failure cases delimit that use.

References

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://aclanthology.org/2021.tacl-1.60/.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.04560.

IUPAC. Periodic table of the elements, 2022. URL https://iupac.org/wp-content/uploads/2022/05/IUPAC_Periodic_Table_A3-04May22.pdf. 4 May 2022. Historical study accessed 7 November 2026.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of EMNLP*, pp. 8849–8861, 2021. doi: 10.18653/v1/2021.emnlp-main.697. URL https://aclanthology.org/2021.emnlp-main.697/.

László Lovász. Operations with structures. *Acta Mathematica Academiae Scientiarum Hungaricae*, 18: 321–328, 1967.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of EMNLP*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Eric Mitchell, Joseph J. Noh, Siyan Li, William S. Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D. Manning. Enhancing self-consistency and performance of pre-trained language models through natural language inference. In *Proceedings of EMNLP*, pp. 1754–1768, 2022. doi: 10.18653/v1/2022.emnlp-main.115. URL https://aclanthology.org/2022.emnlp-main.115/.

Wuming Pan. Token Space: A category theory framework for AI computations, 2026. URL https://arxiv.org/abs/2404.11624v4. arXiv:2404.11624v4.

U.S. Census Bureau. County adjacency file, 2023. URL https://www.census.gov/geographies/reference-files/2023/geo/county-adjacency.html. Accessed 8 November 2026.

U.S. Census Bureau. County adjacency file, 2024. URL https://www.census.gov/geographies/reference-files/2024/geo/county-adjacency.html. Accessed 8 November 2026.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://arxiv.org/abs/2305.10037.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2203.11171.