arXiv:2607.17219v1 [cs.CL] 19 Jul 2026

Auditing Question-Order Effects in Large Language Models with the QQ Equality: Mechanism Characterization and a Saturation Caveat

Pilsung Kang
Department of Software Science, Dankook University
Yongin, South Korea
pilsungk@dankook.ac.kr

Abstract

Human survey respondents exhibit question-order effects that satisfy the QQ (quantum question) equality, an a priori, parameter-free prediction of the projective quantum question-order model. We develop the QQ equality into an audit criterion for sequential judgments of autoregressive large language models (LLMs). Theoretically, we characterize which mechanism classes satisfy it robustly: marginal-independent kernels satisfy QQ iff all four mismatch transition rates coincide (a class containing the 2D rank-1 projective model with a fixed measurement pair under state variation); a polarity- and position-dependent repetition family is characterized by an exact cross-symmetry condition with closed-form violations; QQ-satisfying behaviors are closed under order-matched mixing; and the rank-2 Contextuality-by-Default criterion translates into audit coordinates as $|q_{\mathrm{QQ}}| \leq \mathrm{OSS}$, where OSS (the order-sensitivity score) totals the order sensitivity of the two marginals. Methodologically, we develop a pre-specified, audit-logged pipeline applicable to any model exposing next-token log-probabilities; it combines worst-case robustness envelopes, sampling-consistency spot checks, full label counterbalancing, and a saturation diagnostic. Empirically, in a first-signal pilot on an open-weight instruction-tuned model under two framings, all pre-specified health gates passed, yet 17/18 and 7/8 item pairs, respectively, were saturated (near-deterministic), and no item was certified residually contextual. Forced-binary next-token log-probabilities were thus inadequate for distribution-level QQ audits under the tested model and prompting conditions; we recommend pre-specified saturation diagnostics whenever next-token distributions are treated as survey-response distributions.

1 Introduction

Question order changes human survey answers, sometimes dramatically. In a Gallup poll conducted in September 1997, half of the respondents were asked whether Bill Clinton was honest and trustworthy and then the same question about Al Gore; the other half answered the same two questions in the opposite order. Asked first, Clinton received a 50% agreement rate and Gore 68% — an 18-point gap; asked second, Clinton rose to 57% while Gore fell to 60%, shrinking the gap to 3 points [1, 2]. Order effects of this kind are pervasive and well documented, and by themselves they are not surprising.

What is surprising is a sharper regularity hiding beneath them. Wang et al. [3] documented that across 72 datasets, mostly national field studies (plus laboratory experiments), the two order-conditioned joint distributions satisfy the *QQ (quantum question) equality*

$$q_{\mathrm{QQ}} := \big[p_{AB}(A_y, B_n) + p_{AB}(A_n, B_y)\big] - \big[p_{BA}(B_y, A_n) + p_{BA}(B_n, A_y)\big] = 0, \tag{1}$$

where p_{AB} and p_{BA} denote the joint answer distributions when question A is asked first and when question B is asked first, respectively, and A_y, A_n, B_y, B_n denote "yes" and "no" answers to each

question. This is an a priori, parameter-free prediction of the projective quantum question-order model [2]. Equivalently, the probability that the two answers *disagree* is the same in both orders: in the Clinton–Gore poll above, the disagreement probability was 0.221 in one order and 0.225 in the other, even though the individual agreement rates moved by up to eight points. The equality is thus a non-parametric constraint on the pair of joint distributions — not a no-order-effect claim — and its empirical success is what makes it a candidate audit criterion rather than a curiosity.

Large language models (LLMs) are increasingly used as proxies for human respondents, and exhibit well-documented order and format sensitivities [4–6]. This raises a natural audit question: does the sequential judgment behavior of an autoregressive LLM satisfy the same non-parametric constraint that human populations do? To the best of our knowledge (within our 2026-07 search scope; Section 2), no prior work has tested the QQ equality on LLMs under a multi-turn sequential protocol with answers committed to the conversation history.

Testing this constraint on an LLM is not a direct transcription of the human protocol, because the measured object differs. A human study estimates response distributions from repeated sampling across respondents; an LLM study typically treats the model's next-token probability distribution, read out at a forced-choice answer position, as the response distribution of a synthetic respondent. This substitution is convenient and widely used, but it is a measurement assumption, not a fact: the constraint is informative only if the measured distribution actually carries distributional structure, and nothing guarantees that an instruction-tuned model exposes such structure at a forced binary choice. Making this assumption explicit, testable, and — where it fails — auditable is as much a part of our task as the QQ test itself.

One caution governs the interpretation of any outcome. We deliberately use compatibility language throughout: QQ satisfaction is consistent with projective measurement *and* with symmetric classical repetition (Proposition 4); QQ violation rejects only the standard projective model, not "quantum-like" accounts in general — quantum instruments based on positive operator-valued measures (POVMs) can violate QQ [7].

Answering this question well therefore requires more than running a test. It requires knowing *what a result would mean* (which mechanism classes are compatible with satisfaction or violation), *whether the measurement instrument is adequate*, and *what an inadequate instrument looks like in practice*. Our contributions track exactly these three needs:

1. **Mechanism characterization (Section 3).** We formalize mechanisms as maps from item parameters to sequential response kernels and characterize robust QQ satisfaction in two families (Theorem 3, Proposition 5), prove closure under order-matched mixing (Proposition 7), exhibit a classical mechanism that satisfies QQ identically while producing order effects (Proposition 4), and translate the rank-2 Contextuality-by-Default (CbD) criterion into audit coordinates (Corollary 6). The result is a discriminant table that turns audit outcomes into compatibility statements.

2. **Audit methodology (Section 4).** A pre-specified, audit-logged pipeline — applicable to any model that exposes next-token log-probabilities and admits single-token answer labels — for the *binary-conditioned audit estimand* $P(y \mid y \text{ or } n)$ under multi-turn sequential presentation, with exact robustness-envelope decisions over unassigned probability mass, pre-specified Clopper–Pearson spot checks, full label counterbalancing with semantic canonicalization, and a saturation diagnostic.

3. **A first-signal pilot and a measurement caveat (Section 5).** On Qwen3-4B-Instruct-2507 under two framings (direct evaluation; survey-respondent persona), all pre-specified health checks pass, yet the response distributions are saturated (near-deterministic) for 17/18 and

7/8 item pairs respectively. Observed QQ violations were dominated by deterministic order flips; no item was certified residually contextual (a few were indeterminate at the certification bounds), and label assignment has a substantive effect. The pilot's central finding is that forced-binary next-token log-probabilities failed to expose the distributional structure the QQ audit needs *under the tested model and prompting conditions* — and that a pre-specified saturation diagnostic catches this early and auditably.

2 Related Work

Our work sits at the intersection of two literatures: the quantum-cognition line that produced the QQ equality and its contextuality-theoretic refinements, and the language-model evaluation line that treats LLMs as survey respondents and scrutinizes how their answers are measured.

QQ equality and quantum question-order models. Wang and Busemeyer [2] derived the QQ equality from projective sequential measurement and showed that standard Bayesian and Markov accounts do not generally satisfy it; Wang et al. [3] confirmed it across 72 datasets. The Supporting Information of [3] further shows that a repeat-choice model constrained to satisfy QQ predicts context effects proportional to marginal differences — empirically rejected in humans ($r = .04$) — and that an anchoring-adjustment model satisfies QQ iff its two order biases are equal; both facts foreshadow our characterization results, which we state and prove at the mechanism level. Ozawa and Khrennikov [8] unify order effects, response replicability, and the QQ equality with quantum instruments; Lebedev and Khrennikov [7] show general POVMs can violate QQ.

Contextuality-by-Default. For inconsistently connected systems, CbD [9] provides the correct contextuality criterion; for rank-2 cyclic systems it reduces to a single inequality that we show equals $|q_{\mathrm{QQ}}| \leq \mathrm{OSS}$ in audit coordinates (Corollary 6). Prior CbD analyses of question-order data interpret human order effects as direct influences. Recent concurrent work applies CbD to generative LLMs in other task families: Kumar et al. [10] report irreducible context-dependence in a pronoun-inference setting, while Tuan [11] find zero corrected CbD degree in an enterprise-agent auditing pilot, with direct influence as the dominant signal — consistent with our findings in the question-order setting.

Quantum-inspired tests on LLMs. Imannezhad et al. [12] test conjunction/disjunction fallacies and binary complementarity on GPT-5; Aerts et al. [13] report Bell-type violations and Bose–Einstein fits in concept combinations (interpretation debated); Lo et al. [14] report large-scale sheaf- and CbD-contextuality in BERT probability distributions over a co-reference-style linguistic schema. None tests the QQ equality under sequential multi-turn presentation.

LLMs as survey respondents. Tjuatja et al. [4] document human-like response biases; Rupprecht et al. [5] test 18 LLMs on World Values Survey items under ten perturbations (334,800 interviews), finding a consistent recency bias in almost all tested models; Sandhan et al. [6] evaluate LLM personality tests with conversational context and report model-dependent sensitivity to question ordering. Our results add a distribution-level caveat to this literature: under forced-binary formats the "response distribution" of an instruction-tuned model can be an artifact of a saturated next-token probability distribution, so distribution-level constructs (including QQ) require measurement designs that restore genuine dispersion.

First-token measurement critique. Wang et al. [15] show first-token probabilities can disagree with generated text answers. Our saturation finding extends the critique from *disagreement* to *degeneracy*: even when first-token measurement is internally consistent (our logprob and trajectory-sampling pathways agree, in the pre-specified spot checks, to exact Clopper–Pearson standards), the measured distribution may carry almost no information beyond its argmax.

3 Theory: Which Mechanisms Satisfy the QQ Equality?

QQ satisfaction alone does not identify a unique generating mechanism, and QQ violation alone does not establish residual contextuality. This section therefore characterizes mechanism classes compatible with satisfaction or violation and translates the rank-2 CbD criterion into audit coordinates, providing the interpretive basis for the discriminant table and empirical analysis.

3.1 Setup

For binary questions A, B with answers $\alpha, \beta \in \{y, n\}$, define order-conditioned kernels $p_A(\alpha)$, $K_{AB}(\beta \,|\, \alpha)$ and $p_B(\beta)$, $K_{BA}(\alpha \,|\, \beta)$, with joints $\text{joint}_{AB}(\alpha, \beta) = p_A(\alpha)K_{AB}(\beta \,|\, \alpha)$ (and symmetrically for BA). The observable behavior space is 6-dimensional. Writing M_{XY} for the probability that the two answers disagree in order XY, $q_{\text{QQ}} = M_{AB} - M_{BA}$.

Definition 1 (Mechanism; robust QQ)**.** A *mechanism* maps item parameters $\theta = (p_A, p_B, \dots)$ to kernel pairs, with fixed internal parameters. It satisfies QQ *robustly* if $q_{\text{QQ}} = 0$ on an admissible domain containing an open subset of $(0, 1)^2$ in the first-position marginals.

Finite experiments certify compatibility only, never robustness. We write the mismatch transition rates as $a = K_{AB}(n|y)$, $b = K_{AB}(y|n)$, $c = K_{BA}(n|y)$, and $d = K_{BA}(y|n)$. Order effects on the marginals are $OE_A = p_{BA}(A_y) - p_{AB}(A_y)$ and $OE_B = p_{AB}(B_y) - p_{BA}(B_y)$, the change in each question's "yes" marginal when it is asked second rather than first. We abbreviate q for q_{QQ} where unambiguous.

Throughout the paper, results are stated and proved in classical probability terms (response kernels and transition rates); quantum models enter only as particular mechanism classes to be included in or excluded from compatibility statements (e.g., Theorem 2), and no quantum formalism is needed to follow the audit methodology or the empirical results.

3.2 Results

Theorem 2 (Wang–Busemeyer [2])**.** Projective sequential measurement satisfies $q_{\text{QQ}} = 0$ for every state and every pair of binary projective measurements.

We reproduce this numerically to machine precision in our validation suite.

Theorem 3 (Marginal-independent kernels)**.** If (a, b, c, d) are constants independent of (p_A, p_B), then with $u = p_A(y)$, $v = p_B(y)$,

$$q_{\text{QQ}} = u(a - b) - v(c - d) + (b - d), \tag{2}$$

and robust QQ holds **iff** $a = b = c = d$.

Proof. Coefficient elimination on an open set. □

The 2D rank-1 projective model **with a fixed pair of measurements under state variation** lies in this class, with $a = b = c = d = \mathrm{Tr}(P_{A,y}P_{B,n}) = |\langle A_y|B_n\rangle|^2$; if projectors vary across items, the fixed-kernel hypothesis of Theorem 3 no longer applies.

Proposition 4 (Symmetric repetition)**.** The kernels $K_{AB}(\beta\,|\,\alpha) = r\,\mathbf{1}[\beta = \alpha] + (1-r)\,p_B(\beta)$ and, symmetrically, $K_{BA}(\alpha\,|\,\beta) = r\,\mathbf{1}[\alpha = \beta] + (1-r)\,p_A(\alpha)$, with a single fixed r, satisfy QQ identically while producing order effects $OE_B = r\,(p_A(y) - p_B(y))$. Moreover its mismatch rates $a = (1-r)\big(1-p_B(y)\big)$, $b = (1-r)\,p_B(y)$ differ whenever $0 \le r < 1$ and $p_B(y) \neq 1/2$, so **at the kernel level** this mechanism is not realizable by any 2D rank-1 projective model at such items (identification of both kernel branches from observed behavior additionally requires interior first-answer marginals). Hence QQ satisfaction does not uniquely identify a (2D rank-1) projective mechanism.

Proposition 5 (Mixture-repetition family)**.** With fixed polarity- and position-dependent retention rates $r_{AB}[\cdot]$, $r_{BA}[\cdot]$,

$$q_{\mathrm{QQ}} = p_A(y)p_B(n)\,(r_{BA}[n] - r_{AB}[y]) + p_A(n)p_B(y)\,(r_{BA}[y] - r_{AB}[n]), \tag{3}$$

so robust QQ holds **iff** $r_{AB}[y] = r_{BA}[n]$ and $r_{AB}[n] = r_{BA}[y]$ (cross-polarity position symmetry). Position-invariant retention gives $q = (r[n] - r[y])\,(p_A(y) - p_B(y))$ (polarity-dependent repetition); polarity-invariant retention gives $q = (r_{BA} - r_{AB})\,S$ with $S := p_A(y)p_B(n) + p_A(n)p_B(y)$ (position-dependent repetition). These closed forms are pre-specified alternative explanations for violations.

Corollary 6 (CbD criterion in audit coordinates)**.** For the rank-2 cyclic system with ± 1 coding, the Kujala–Dzhafarov noncontextuality criterion [9] is equivalent to

$$|q_{\mathrm{QQ}}| \;\le\; \mathrm{OSS} := |OE_A| + |OE_B|, \tag{4}$$

where we call OSS the *order-sensitivity score* of the item. Thus $q_{\mathrm{QQ}} = 0$ implies noncontextuality; violations with $|q| \le \mathrm{OSS}$ are compatible with a noncontextual coupling that allows direct influences (the criterion does not identify direct influence as the generating cause); only $|q| > \mathrm{OSS}$ certifies residual contextuality. OSS is thereby an inconsistent-connectedness allowance. (The mathematical content is [9]; our contribution is the translation identity $|E[AB]_{AB} - E[AB]_{BA}| = 2|q_{\mathrm{QQ}}|$, $\Delta = 2\,\mathrm{OSS}$.)

Proposition 7 (Order-matched mixing closure)**.** q_{QQ} is a linear functional of the joint pair, so any mixture of QQ-satisfying behaviors with weights identical across the two orders (possibly θ-dependent) satisfies QQ; order-dependent component selection **can** violate QQ (a possible violation channel, not a necessity). This provides one sufficient route by which heterogeneous populations may preserve QQ satisfaction in aggregate, is consistent with the pooled estimand used for counterbalanced label mappings (Section 4), and marks order-dependent routing (e.g., in mixture-of-experts decoding) as an audit checkpoint.

Two numerical facts are retained as conjecture-generating evidence: symmetric unsharp qubit Lüders instruments preserved QQ in all tested configurations, and violations were observed in some tested asymmetric variants (consistent with the possibility results of [7]). Open problems include the general characterization and representation question, Bayesian-update boundaries, and higher-rank cyclic audits.

Observed pattern (tested scope)	Compatible mechanism classes
broad empirical QQ satisfaction	marginal-independent (a=b=c=d); projective; symmetric/cross-symmetric repetition; order-matched mixtures
violations fitting $q \sim (r_n - r_y)(p_A(y) - p_B(y))$	polarity-dependent repetition
violations fitting $q \sim (r_{BA} - r_{AB})S$	position-dependent repetition
both closed forms rejected, Γ envelope ≤ 0	mechanisms compatible with noncontextual couplings allowing direct influences; candidate (not established) examples: tested asymmetric Lüders variants, order-dependent mixing
Γ envelope > 0	certified residual (CbD) contextuality

Table 1: Discriminant table mapping audit outcomes to compatibility statements. Here $\Gamma := |q| - \mathrm{OSS}$, so that $\Gamma \leq 0$ is the noncontextuality condition of Corollary 6; certified envelope bounds on Γ are constructed in Section 4.

3.3 Discriminant table

Table 1 assembles the results of this section into an interpretation aid for audit outcomes. Each observed pattern is mapped to the mechanism classes *compatible* with it — by Proposition 4, no row licenses a unique mechanism identification. The first row collects the satisfaction classes characterized by Theorem 3 and Propositions 4–5; the two closed-form rows are the pre-specified alternative explanations of Proposition 5; and the final two rows apply the certification logic of Corollary 6. Section 5 uses this table to interpret the pilot results.

4 Audit Methodology

This section translates the theoretical distinctions above into a reproducible audit pipeline for sequential judgments of LLMs. The procedure defines the estimand and history-conditioned branching protocol, controls label and unassigned-mass effects, and checks both computational consistency and distributional adequacy before QQ or contextuality verdicts are interpreted.

Estimand. The *binary-conditioned audit estimand* is $P(y \mid y \text{ or } n) = e^{\ell_y}/(e^{\ell_y} + e^{\ell_n})$, computed from the next-token log-probabilities of the canonical single-token labels. Joint probabilities are assembled as $p(A_a, B_b) = P(A_a)\, P(B_b \mid \text{history contains } A_a)$ under a history-conditioned forced-branch sequential protocol: the first question is asked, each candidate answer is committed to the conversation history in turn, and the second question is asked conditional on that history, in both orders. This estimand differs from full-vocabulary generation; the gap is tracked through the recorded raw label mass at every position.

Label counterbalancing. Original survey wordings and response options are preserved; options map to neutral tokens A/B under *both* assignments (map-1, map-2), each applied identically across orders and protocols. Results are canonicalized back to semantic options before pooling with equal weights. Pooling is justified by the linearity of q_{QQ} in the joint pair together with the pre-specified equal-weight estimand; Proposition 7 additionally clarifies how such order-matched mixtures preserve QQ structure. Per-mapping results are reported as label-effect diagnostics (the two mappings are experimental conditions, not latent population components).

Decision rules (pre-specified; frozen before execution and audit-logged). Unassigned probability mass ("other") is handled by a worst-case robustness envelope: q_{QQ} is multilinear in the six mass-allocation variables, so its exact range over the allocation box is attained on 2^6 vertices; the equivalence decision (SATISFIED/VIOLATED/INDETERMINATE at $\varepsilon_{QQ} = 0.02$, with a floating-point tolerance) is taken over the envelope. The margin ε_{QQ} was fixed a priori as a practical-equivalence band wide enough to cover approximate satisfaction by partial-revision processes, for which simulations in the Supporting Information of Wang et al. [3] report $|q| < 0.015$. $\Gamma = |q| - \mathrm{OSS}$ is only piecewise multilinear, so certified bounds are assembled from exact component envelopes (sound by elementary inequalities; conservatism verified empirically). Logprob and trajectory-level rejection sampling are cross-checked by exact Clopper–Pearson intervals with Bonferroni correction (per item-order, $\alpha = 0.01$), a method selected after an observed simultaneous undercoverage of Wilson intervals (0.980 vs 0.995 at $n = 200$ in calibration).

Saturation diagnostic (pre-specified). An item is *saturated* if more than half of its measured positions have $\max(p, 1-p) > 0.99$. Saturated items carry little distributional information; their QQ verdicts are reported but weakly interpretable.

Health gates. Gate G1 requires the canonical labels to be single tokens in the rendered chat-template context. Gate G2 requires the other mass to stay below 10^{-2} at every position, screening gross mass leakage; the decision-relevant quantity is the resulting envelope width, which is reported with the results. Gate G3 cross-checks logprob against sampling as above, run as pre-specified spot checks rather than exhaustively: under Track M, items s1-01 and s3-01, both orders, $n_{\mathrm{rej}} = 200$ accepted joint samples each; under Track P, item s1-01, both orders, $n_{\mathrm{rej}} = 200$.

The mathematical core was validated against ground-truth agents (projective, repetition families, Lüders instruments), recovering $q = 0$ at machine precision and all closed-form violations exactly; the full software assertion suite ships with the code as part of the reproducibility materials.

5 Empirical Pilot

This section applies the audit pipeline to a first-signal pilot on an instruction-tuned LLM, evaluating both the technical reliability of the procedure and whether the resulting response distributions are adequate for meaningful QQ and CbD analysis.

Setup. All measurements were taken on Qwen3-4B-Instruct-2507 [16], run in bfloat16 (bf16) in non-thinking mode; this is a single-model study, and no cross-model generalization is claimed (see Limitations). The item set consists of 18 question pairs: 10 historical survey pairs with original wordings and response options, drawn from Wang et al. [3] and related classic survey sources (Gallup, Pew, SRC) and forming stratum S1; 3 name-swapped memorization controls (S2); and 5 novel judgment pairs answered with Yes/No labels (S3). Two framings were tested, labeled Track M and Track P, keeping the names used in the pre-specified protocol and its audit records. Track M applies a direct evaluation instruction; Track P, run once as a redesign rerun frozen before execution on the 8 year-fixed S1/S2 pairs, applies a survey-respondent persona with a per-item reference year. We use M and P as shorthand in tables and figures; the full prompts and the item manifest are given in Appendix B. Measurement was deterministic at the logprob level, with decoding for the sampling checks fixed at temperature 1 and top-p 1. Every run writes an immutable audit record containing the model and tokenizer revisions, the chat-template hash, label token ids, mapping assignment, seeds, and rejection attempt and accept counts.

Gate results. Gates G1 and G2 passed in every run, and gate G3 passed in all pre-specified spot checks: the maximum other mass was $\sim 2 \times 10^{-7}$, and the consistency checks passed without retries. The implementation and its internal-consistency checks are healthy; what follows concerns the adequacy of the model–prompt–estimand combination for the target construct, not implementation health.

Results. Table 2 reports the pooled Track M verdicts for all 18 items; Table 3 compares the matched eight items across the two framings; Table 4 tabulates the per-mapping label-effect diagnostics; Figure 1 visualizes the position-level saturation, and Figure 2 the label effect. Four findings follow.

id	OSS	q_{QQ}	QQ	Γ class	sat
s1-01	0.742	+0.182	VIOLATED	NONCONTEXTUAL	Y
s1-02	0.981	+0.981	VIOLATED	**INDETERMINATE**	Y
s1-04	0.406	+0.002	SATISFIED	NONCONTEXTUAL	Y
s1-05	0.733	−0.730	VIOLATED	NONCONTEXTUAL	Y
s1-06	0.009	−0.006	SATISFIED	NONCONTEXTUAL	Y
s1-07	0.774	−0.000	SATISFIED	NONCONTEXTUAL	Y
s1-09	0.309	−0.309	VIOLATED	NONCONTEXTUAL	Y
s1-10	0.025	−0.023	VIOLATED	NONCONTEXTUAL	Y
s1-11	0.741	−0.049	VIOLATED	NONCONTEXTUAL	Y
s1-12	0.006	+0.006	SATISFIED	NONCONTEXTUAL	Y
s2-01	0.804	+0.182	VIOLATED	NONCONTEXTUAL	Y
s2-02	0.000	+0.000	SATISFIED	NONCONTEXTUAL	Y
s2-09	0.595	+0.468	VIOLATED	NONCONTEXTUAL	Y
s3-01	0.562	+0.562	VIOLATED	**INDETERMINATE**	Y
s3-03	0.500	+0.500	VIOLATED	NONCONTEXTUAL	Y
s3-08	1.369	−0.614	VIOLATED	NONCONTEXTUAL	Y
s3-09	0.775	+0.687	VIOLATED	NONCONTEXTUAL	Y
s3-11	0.959	−0.890	VIOLATED	NONCONTEXTUAL	N

Table 2: Track M pooled results (18 items). The sat column marks item-level saturation (Y/N) under the pre-specified rule of Section 4. S1/S2 use A/B-mapped original options; S3 uses Yes/No — metrics are not compared numerically across strata. Envelope intervals are indistinguishable from point values at the displayed precision (other mass $\sim 10^{-7}$). Classifications use full-precision certified bounds; displayed values are rounded.

1. *Saturation dominates, and persists under persona framing.* Under Track M, 17 of the 18 items were saturated. The persona framing moved individual values dramatically — s1-06's q, for example, shifted from -0.01 to -1.00, the theoretical extremum — but it did not open the distributions: 7 of 8 items remained saturated, with only s1-11 de-saturating (Figure 1). We therefore report that saturation *persisted* under respondent-persona framing within the tested conditions, rather than claiming strict invariance, since one item de-saturated and values moved substantially. The verdict is insensitive to the threshold choice: at thresholds 0.95, 0.975, and 0.99 the saturated-item counts are 18/18, 18/18, and 17/18 under Track M, and 7/8 throughout under Track P.

2. *Deterministic patterns and label effects dominate the QQ verdicts.* The pooled Track M verdicts are 13 VIOLATED, 5 SATISFIED, and 0 INDETERMINATE, and the items fall into

three families: order-invariant deterministic items (trivially SATISFIED with $|q| \approx \text{OSS} \approx 0$), order-flipping deterministic items (VIOLATED with $|q| \approx \text{OSS}$), and label-dominated items. Crucially, there was **no certified residual contextuality**: no item had certified $\Gamma_{\text{lower}} > 0$ in either framing. Under Track M, 16 items were certified noncontextual and 2 (s1-02, s3-01) were indeterminate at the certification bounds (Table 2); under Track P, 5 were certified noncontextual and 3 (s1-02, s1-12, s2-01) indeterminate (Table 3). The higher indeterminate share under Track P reflects items sitting exactly on the $|q| = \text{OSS}$ certification boundary: order-flipping items with $|q| \approx \text{OSS} \approx 1$ (s1-02, s2-01) and a near-null item with $|q| \approx \text{OSS} \approx 0$ (s1-12) alike.

3. *Label assignment has a substantive effect.* Across the counterbalanced mappings, per-item q_{QQ} differed by up to 1.48 (s1-01) and OSS by up to 1.52 (s2-01); in the saturated regime the A/B token choice itself can flip answers (Table 4, Figure 2). Counterbalancing with semantic canonicalization is therefore not optional for audits of this kind, consistent with the label and position biases reported by Rupprecht et al. [5].

4. *The memorization alert was not triggered.* The pre-specified rule raises an alert if $|\text{OSS}_{S1} - \text{OSS}_{S2}| > 0.3$ on at least 2 of the 3 matched S1/S2 pairs; this occurred on only 1 of 3 (s1-02/s2-02), so the alert was not triggered. This check is descriptive; comparison against human joint benchmarks remains pending.

id	year	sat M→P	q_{QQ} M→P	OSS M→P	QQ (P)	Γ (P)
s1-01	1997	Y→Y	$+0.18 \to -0.57$	$0.74 \to 0.57$	VIOLATED	NC
s1-02	1995	Y→Y	$+0.98 \to +0.11$	$0.98 \to 0.11$	VIOLATED	**IND**
s1-04	1996	Y→Y	$+0.00 \to +0.65$	$0.41 \to 0.65$	VIOLATED	NC
s1-06	1979	Y→Y	$-0.01 \to -1.00$	$0.01 \to 1.00$	VIOLATED	NC
s1-11	2008	Y→N	$-0.05 \to -0.24$	$0.74 \to 0.89$	VIOLATED	NC
s1-12	2012	Y→Y	$+0.01 \to +0.00$	$0.01 \to 0.00$	SATISFIED	**IND**
s2-01	1997	Y→Y	$+0.18 \to -1.00$	$0.80 \to 1.00$	VIOLATED	**IND**
s2-02	1995	Y→Y	$+0.00 \to +0.01$	$0.00 \to 0.47$	SATISFIED	NC

Table 3: Track M vs Track P, 8 year-fixed pairs. Under Track P no item had certified $\Gamma_{\text{lower}} > 0$ (5 certified noncontextual (NC), 3 indeterminate (IND) at the certification bounds). Classifications use full-precision certified bounds; displayed values are rounded.

Main finding (methodological). Under the tested instruction-tuned model and prompting conditions, forced-binary next-token log-probabilities (i) *revealed substantial order and label sensitivity* — the deterministic flips and label-mapping effects above are observed properties of the binary-conditioned estimand; but were (ii) *inadequate for measuring population-like dispersed response distributions*, since the binary-conditioned distribution was nearly deterministic under both framings; and consequently (iii) *only weakly informative for fine-grained mechanism discrimination and residual-contextuality detection*: boundary-dominated behavior ($|q| \approx \text{OSS}$, including the near-null case $|q| \approx \text{OSS} \approx 0$) limited what the discriminant machinery of Section 3 could resolve. As a general (not universally proven) methodological warning, we recommend a pre-specified saturation diagnostic as a standard health check for any study treating LLM next-token distributions as survey-response distributions.

id	q (map-1)	q (map-2)	$\lvert\Delta q\rvert$	OSS (map-1)	OSS (map-2)	$\lvert\Delta\text{OSS}\rvert$
s1-01	−0.560	+0.924	1.484	0.560	0.924	0.364
s1-02	+0.963	+0.999	0.036	0.963	0.999	0.036
s1-04	+0.000	+0.003	0.003	0.006	0.818	0.812
s1-05	−0.962	−0.498	0.465	0.963	0.503	0.461
s1-06	−0.014	+0.002	0.017	0.014	0.004	0.010
s1-07	−0.000	+0.000	0.000	1.358	0.189	1.169
s1-09	−0.618	+0.000	0.619	0.618	0.004	0.614
s1-10	−0.046	−0.000	0.045	0.049	0.000	0.049
s1-11	−0.098	+0.001	0.099	1.456	0.026	1.431
s1-12	+0.000	+0.011	0.011	0.000	0.011	0.011
s2-01	+0.363	−0.000	0.364	1.562	0.046	1.516
s2-02	+0.000	+0.000	0.000	0.001	0.000	0.001
s2-09	−0.024	+0.959	0.983	0.215	0.976	0.762

Table 4: Label-effect diagnostics (A/B items, Track M): q_{QQ} and OSS by mapping. All columns are rounded independently from full-precision values, so displayed differences may deviate from differences of displayed values in the last digit.

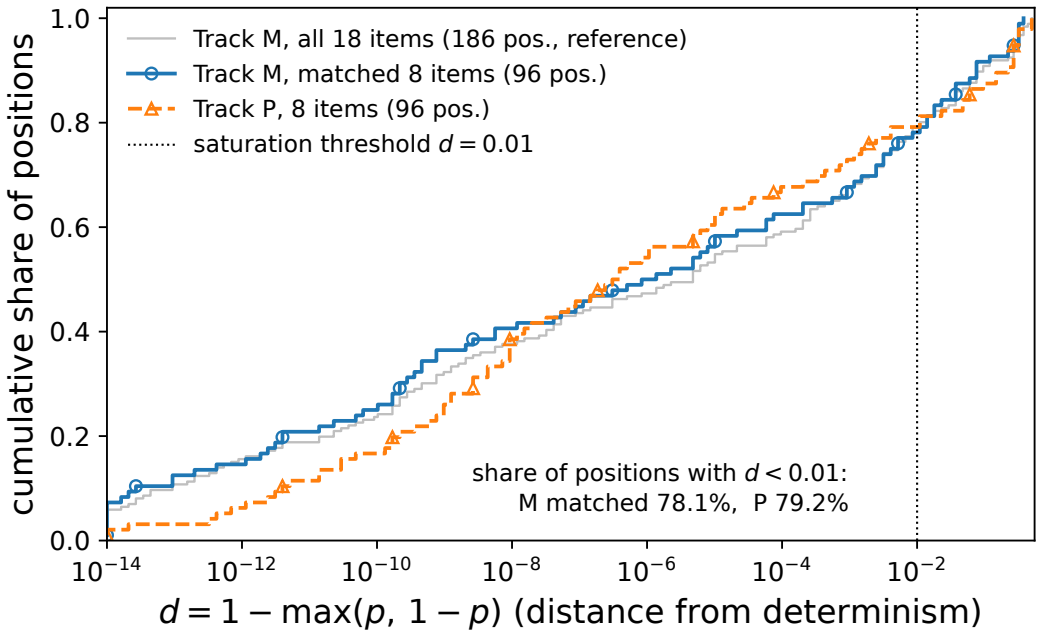

Figure 1: Empirical cumulative distribution function (CDF) of the distance from determinism $d = 1 - \max(p, 1-p)$, log scale (d equals the probability of the minority answer, i.e., the total-variation distance to the nearest deterministic distribution). The primary comparison is between the matched eight-item subset under Track M and Track P (96 positions each); the full Track M set is shown as a reference curve. On the matched subset, a similar share of positions lies below the $d = 0.01$ saturation threshold (78.1% vs 79.2%), indicating that the respondent-persona framing did not substantially alleviate saturation, although the distributions differ at intermediate scales. Positions with $d = 0$ are clipped to 10^{-14} for display.

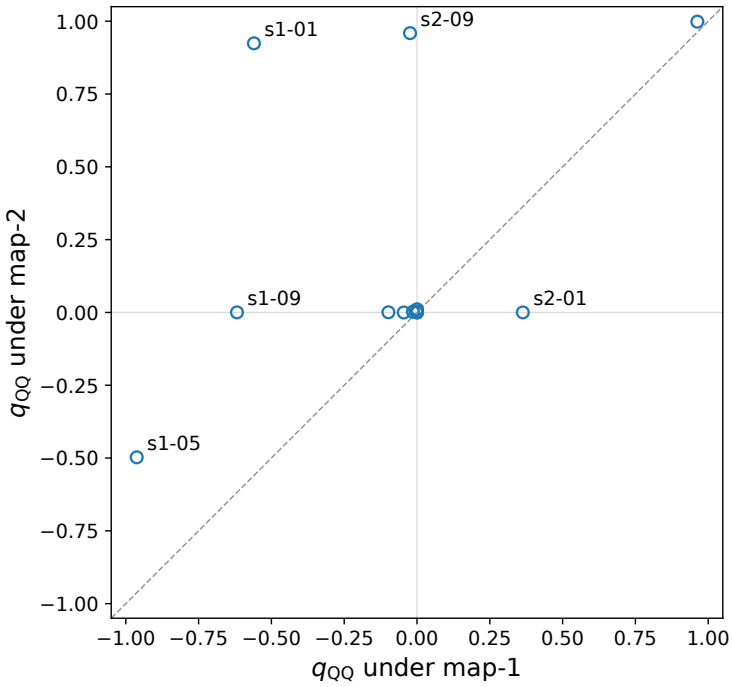

Figure 2: Label effect under counterbalancing: per-item q_{QQ} under map-1 vs map-2 (13 A/B items, Track M). Points on the dashed diagonal indicate mapping-invariant behavior; annotated items deviate by more than 0.3, showing that in the saturated regime the A/B token assignment has a substantive effect on the measured q_{QQ}.

6 Discussion

This section interprets the pilot findings in relation to LLM-as-respondent research, clarifies what they imply for QQ and CbD audits, and outlines measurement designs that may yield more informative response distributions.

Relation to LLM-as-respondent studies. Studies that aggregate LLM "response distributions" from forced-choice next-token measurements implicitly assume dispersion. Our results show that for a modern instruction-tuned model this assumption can fail for nearly all tested items at the single-policy level, with apparent distributional effects driven by deterministic flips and label-assignment effects. This sharpens, at the distribution level, the first-token critique of Wang et al. [15].

What a QQ audit needs instead. The theory clarifies both the opportunity and a hazard. Human aggregates are mixtures of individuals; *if* individual-level behavior satisfies QQ, Proposition 7 explains aggregate satisfaction under order-matched mixing — but the converse direction matters for design: mixture populations (personas; temperature ensembles with order-invariant selection) and alternative measurement designs (base checkpoints, sampling-based estimands) may restore genuine dispersion; mixtures in particular can also average away component-level violations, so a mixture-based audit must report component-level diagnostics alongside aggregate verdicts.

What the pilot does and does not show. This pilot documents a measurement caveat within a deliberately narrow empirical scope. Its evidence comes from one instruction-tuned model

and 18 item pairs, many drawn from U.S. political and social survey contexts, under a single binary-conditioned next-token estimand and two prompt framings. Accordingly, the persistence of saturation should be interpreted as a property of the tested model–prompt–item configuration, not as evidence that instruction-tuned LLMs generally produce saturated response distributions. What the pilot does show is that, under these conditions, a technically healthy and internally consistent audit can still be limited by near-deterministic outputs; it does not establish whether the same limitation persists across model scales, base checkpoints, languages, cultures, or broader item domains.

In addition, the pilot does not show that LLMs are, or are not, "quantum-like". QQ satisfaction is compatible with classical repetition (Proposition 4), QQ violation is compatible with noncontextual direct influences (Corollary 6), and in the saturated regime many verdicts are dominated by deterministic or label-sensitive behavior; the informative content of the pilot therefore concerns the measurement layer more than the mechanism layer. The absence of certified contextuality here is consistent with the concurrent enterprise-agent audit of Tuan [11], where direct influence dominated the signal, but contrasts with the irreducible context-dependence reported by Kumar et al. [10] in a pronoun-inference setting. Task family, estimand construction, and the dispersion of the underlying distributions all differ across these studies, and reconciling the two regimes is a natural target for the distribution-restoring designs above.

For the audit engine. The components that remain useful beyond this pilot are the characterization theorems, the discriminant table, and the validated pipeline with its health gates; these contributions do not depend on the particular LLM tested here. In an audit workflow, the saturation diagnostic serves as a screening check for whether a system's judgment interface exposes sufficient distributional structure for downstream analysis.

7 Limitations

The empirical scope of this study is deliberately narrow. The reported results come from a single instruction-tuned model, with Track M covering 18 English-language item pairs and Track P covering the eight year-fixed S1/S2 pairs under a respondent-persona framing. They therefore do not support cross-model, cross-linguistic, or cross-cultural generalization. Many items were drawn from U.S. political and social survey traditions, although the set also included matched name-substitution controls and newly constructed judgment pairs. We further examined only one measurement scheme: binary-conditioned next-token log-probabilities obtained through single-token forced-choice labels. The use of neutral A/B labels with an explicit per-turn legend is itself part of this design, and other response formats, model scales, base checkpoints, or estimands may produce different patterns of label sensitivity, dispersion, and saturation. Accordingly, our empirical conclusion is limited to the persistence of saturation under the tested model–prompt–item–estimand configurations, rather than to instruction-tuned LLMs or next-token measurement in general.

The inferential and interpretive scope is also limited. Two Track M items and three Track P items remained Γ-indeterminate at the certification bounds. Our envelope procedure certifies noncontextuality when the certified upper bound on Γ is nonpositive. For the indeterminate items, however, failure to certify residual contextuality should not be interpreted as evidence that it is absent. Human joint-probability benchmarks were unavailable for direct calibration, so comparisons with historical survey findings remain qualitative. More fundamentally, QQ satisfaction establishes compatibility with multiple mechanism classes rather than identifying a unique generating mechanism (Proposition 4). Conversely, a QQ violation shows that the measured binary-conditioned joint

distributions are incompatible with the standard projective question-order model; it does not identify the model's internal mechanism or exclude more general quantum-like instrument models [7]. Finally, the claim that no prior study has tested the QQ equality on LLMs under a committed, multi-turn sequential protocol is limited to the literature identified within our July 2026 search window.

8 Conclusion

We developed the QQ equality—an a priori constraint of the standard projective quantum question-order model that has been observed across human survey data—into an audit criterion for sequential binary judgments of autoregressive LLMs. By characterizing robust QQ satisfaction in several mechanism families, establishing closure under order-matched mixing, and translating the rank-2 CbD criterion into the audit inequality $|q_{\mathrm{QQ}}| \leq \mathrm{OSS}$, we clarified what inferential claims QQ satisfaction and violation can and cannot support.

In the empirical pilot, all pre-specified health gates passed, indicating that the sequential branching, label-token measurement, robustness envelopes, and sampling spot checks operated as intended. Nevertheless, the binary-conditioned distributions were saturated for 17 of 18 Track M items and 7 of 8 Track P items, while counterbalanced mappings showed that neutral-label assignment could materially change the measured q_{QQ} and OSS. The pooled audit produced both QQ satisfaction and violation, but no item had a certified $\Gamma_{\mathrm{lower}} > 0$; many violations instead lay near the boundary $|q_{\mathrm{QQ}}| = \mathrm{OSS}$ and were therefore compatible with noncontextual couplings that allow between-order marginal changes, with several items remaining indeterminate at the certification bounds. These findings do not establish whether the model is or is not "quantum-like." Rather, they show that, under the tested model–prompt–item conditions, a technically reproducible forced-binary next-token estimand can remain too nearly deterministic to support fine-grained mechanism discrimination.

Methodologically, the pilot shows that distribution-level audits of LLMs should assess the adequacy of the response distribution before interpreting higher-level structural criteria. Saturation diagnostics, label counterbalancing, and explicit tracking of unassigned probability mass should therefore accompany any use of next-token probabilities as survey-response distributions. Future work should test whether this measurement caveat persists across broader model families and alternative measurement designs, while retaining component-level diagnostics whenever mixture-based approaches are used to restore dispersion.

Code and Data Availability

All experiment code, item sets, the validated mathematical core with its assertion suite, and the immutable audit records (JSON) and run reports for every execution reported here are available at `https://github.com/pilsungk/llm-qq` (commit `4563023`).

Acknowledgments

This work was supported by the National Research Foundation of Korea grant funded by the Korea government (Ministry of Science and ICT), grant number RS-2026-25477171.

References

[1] David W. Moore. Measuring New Types of Question-Order Effects: Additive and Subtractive. *Public Opinion Quarterly*, 66(1):80–91, 2002. doi: 10.1086/338631.

[2] Zheng Wang and Jerome R. Busemeyer. A Quantum Question Order Model Supported by Empirical Tests of an A Priori and Precise Prediction. *Topics in Cognitive Science*, 5(4): 689–710, 2013. doi: 10.1111/tops.12040. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12040.

[3] Zheng Wang, Tyler Solloway, Richard M. Shiffrin, and Jerome R. Busemeyer. Context Effects Produced by Question Orders Reveal Quantum Nature of Human Judgments. *Proceedings of the National Academy of Sciences*, 111(26):9431–9436, 2014. doi: 10.1073/pnas.1407756111.

[4] Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024. doi: 10.1162/tacl_a_00685. URL https://aclanthology.org/2024.tacl-1.56/.

[5] Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. Prompt Perturbations Reveal Human-Like Biases in Large Language Model Survey Responses. In *Proceedings of the Seventh Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 1–21, San Diego, CA, 2026. Association for Computational Linguistics.

[6] Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. CAPE: Context-Aware Personality Evaluation Framework for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10648–10662. Association for Computational Linguistics, 2025.

[7] Aleksandr Lebedev and Andrei Khrennikov. *Quantum-like Modeling of the Order Effect in Decision Making: POVM Viewpoint on the Wang–Busemeyer QQ-equality*, pages 123–128. World Scientific, Singapore, 2023. doi: 10.1142/9789811275999_0010.

[8] Masanao Ozawa and Andrei Khrennikov. Modeling Combination of Question Order Effect, Response Replicability Effect, and QQ-equality with Quantum Instruments. *Journal of Mathematical Psychology*, 100:102491, 2021. ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2020.102491. URL https://www.sciencedirect.com/science/article/pii/S0022249620301152.

[9] Janne V. Kujala and Ehtibar N. Dzhafarov. Proof of a Conjecture on Contextuality in Cyclic Systems with Binary Variables. *Foundations of Physics*, 46(3):282–299, Mar 2016. ISSN 1572-9516. doi: 10.1007/s10701-015-9964-8. URL https://doi.org/10.1007/s10701-015-9964-8.

[10] Sagar Kumar, Ariel Flint, Luca Maria Aiello, and Andrea Baronchelli. Failure of Contextual Invariance in Large Language Models, 2026. URL https://arxiv.org/abs/2603.23485.

[11] Thanh Luong Tuan. Ontology-Amplified Distillation and Contextuality Auditing for Sovereign Enterprise Language Models: A Combined Proof-of-Mechanism and Negative-Results Method Study, 2026. URL https://arxiv.org/abs/2607.11948.

[12] Pegah Imannezhad, Emmanuel M. Pothos, and Andy J. Wills. Divergent Patterns of Probabilistic Reasoning in Humans and GPT-5. *Frontiers in Psychology*, 17:1782184, 2026. ISSN

1664-1078. doi: 10.3389/fpsyg.2026.1782184. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2026.1782184.

[13] Diederik Aerts, Jonito Aerts Arguëlles, Lester Beltran, Suzette Geriente, Roberto Leporini, Massimiliano Sassoli de Bianchi, and Sandro Sozzo. Identifying Quantum Structure in AI Language: Evidence for Evolutionary Convergence of Human and Artificial Cognition. *Entropy*, 28(6):622, June 2026. ISSN 1099-4300. doi: 10.3390/e28060622. URL http://dx.doi.org/10.3390/e28060622.

[14] Kin Ian Lo, Mehrnoosh Sadrzadeh, and Shane Mansfield. Quantum-like Contextuality in Large Language Models. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 481(2319):20240399, 08 2025. ISSN 1364-5021. doi: 10.1098/rspa.2024.0399. URL https://doi.org/10.1098/rspa.2024.0399.

[15] Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416. Association for Computational Linguistics, 2024.

[16] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025.

A Proofs

Throughout, answers are denoted $\alpha, \beta \in \{y, n\}$, and we write $u = p_A(y)$, $v = p_B(y)$. Recall the mismatch transition rates $a = K_{AB}(n|y)$, $b = K_{AB}(y|n)$, $c = K_{BA}(n|y)$, $d = K_{BA}(y|n)$. The disagreement probabilities take the form

$$M_{AB} \;=\; u\,a + (1-u)\,b, \qquad M_{BA} \;=\; v\,c + (1-v)\,d, \tag{5}$$

since in order AB a disagreement occurs when the first answer is y (probability u) and the second flips (rate a), or the first is n (probability $1-u$) and the second flips (rate b); symmetrically for BA. All proofs below are elementary manipulations of (5).

Proof of Theorem 3. From (5),

$$q_{\mathrm{QQ}} = M_{AB} - M_{BA} = u(a-b) - v(c-d) + (b-d), \tag{6}$$

which is affine in (u, v) with coefficients independent of (u, v) by hypothesis. If q_{QQ} vanishes on a set containing an open subset of $(0,1)^2$, the affine function vanishes identically there, so its coefficients vanish: $a = b$, $c = d$, and $b = d$; hence $a = b = c = d$. Conversely, if $a = b = c = d$ then $M_{AB} = a = M_{BA}$ for all (u, v), so $q_{\mathrm{QQ}} \equiv 0$. □

Projective transition symmetry and family-level inclusion. For binary rank-1 projective measurements on a qubit, the second-answer flip rates are transition probabilities between eigenstates: $K_{AB}(n|y) = |\langle B_n|A_y\rangle|^2$ and $K_{AB}(y|n) = |\langle B_y|A_n\rangle|^2$. In dimension 2, orthocomplementarity gives $|\langle B_y|A_n\rangle|^2 = |\langle B_n|A_y\rangle|^2$, and the same quantity appears in both orders since the modulus is symmetric in its arguments. Hence *at any single item*, every 2D rank-1 projective sequential kernel satisfies $a = b = c = d = |\langle A_y|B_n\rangle|^2$, whatever projectors are chosen for that item. If, in addition, one measurement pair is fixed across items (state variation only), (a, b, c, d) are constants of the mechanism and the model lies in the marginal-independent class of Theorem 3; if projectors vary across items, the family-level fixed-kernel hypothesis fails, but the item-level symmetry $a = b$ remains.

Proof of Proposition 4. With $K_{AB}(\beta \,|\, \alpha) = r\,\mathbf{1}[\beta = \alpha] + (1-r)\,p_B(\beta)$, the flip rates are $a = (1-r)\,p_B(n) = (1-r)(1-v)$ and $b = (1-r)\,v$, so by (5)

$$M_{AB} = (1-r)\big[u(1-v) + (1-u)v\big]. \tag{7}$$

The same computation for $K_{BA}(\alpha \,|\, \beta) = r\,\mathbf{1}[\alpha = \beta] + (1-r)\,p_A(\alpha)$ gives $M_{BA} = (1-r)\big[v(1-u) + (1-v)u\big] = M_{AB}$. Hence $q_{\mathrm{QQ}} \equiv 0$ for every (u, v) and every r.

For the order effect, the second-position marginal in order AB is

$$p_{AB}(B_y) = u\big[r + (1-r)v\big] + (1-u)(1-r)v = r\,u + (1-r)\,v, \tag{8}$$

while $p_{BA}(B_y) = v$; by the sign convention of Section 3, $OE_B = p_{AB}(B_y) - p_{BA}(B_y) = r\,(u-v)$.

Finally, $a - b = (1-r)(1-2v) \neq 0$ whenever $0 \leq r < 1$ and $v \neq 1/2$. By the projective transition symmetry above, *every* 2D rank-1 projective sequential kernel has $a = b$ at any single item, whatever projectors are chosen for that item; hence at items with $v \neq 1/2$ the repetition kernel is not realizable by any such projective kernel — an item-level exclusion that does not require the fixed-pair hypothesis. Identifying both kernel branches from observed behavior requires both first answers to occur, i.e., interior first-answer marginals $0 < u, v < 1$. □

Proof of Proposition 5. With retention rates depending on polarity and position, $K_{AB}(n|y) = (1 - r_{AB}[y])\, p_B(n)$ and $K_{AB}(y|n) = (1 - r_{AB}[n])\, p_B(y)$, so

$$M_{AB} = (1 - r_{AB}[y])\, u(1 - v) + (1 - r_{AB}[n])\, (1 - u)v, \tag{9}$$

$$M_{BA} = (1 - r_{BA}[y])\, v(1 - u) + (1 - r_{BA}[n])\, (1 - v)u, \tag{10}$$

and subtracting,

$$q_{\mathrm{QQ}} = u(1 - v)\, \big(r_{BA}[n] - r_{AB}[y]\big) + (1 - u)v\, \big(r_{BA}[y] - r_{AB}[n]\big). \tag{11}$$

The two monomials $u(1 - v)$ and $(1 - u)v$ are linearly independent as functions on any open subset of $(0, 1)^2$, so (11) vanishes on such a set iff both coefficients vanish: $r_{AB}[y] = r_{BA}[n]$ and $r_{AB}[n] = r_{BA}[y]$. The "iff" is relative to the robust (open-domain) notion of Definition 1: on a boundary segment or a restricted item family the two monomials can be linearly dependent, and (11) may vanish without cross-polarity symmetry. The two special cases follow by substitution: position-invariant retention ($r_{AB}[\cdot] = r_{BA}[\cdot] = r[\cdot]$) gives $q_{\mathrm{QQ}} = (r[n] - r[y])(u - v)$, and polarity-invariant retention ($r_{XY}[y] = r_{XY}[n] = r_{XY}$) gives $q_{\mathrm{QQ}} = (r_{BA} - r_{AB})\, S$ with $S = u(1 - v) + (1 - u)v$. □

Proof of Corollary 6. Code answers as ± 1 ($y \mapsto +1$, $n \mapsto -1$). Following Contextuality-by-Default, the two orders are two contexts, and the variables measured in them are distinct random variables A^{AB}, B^{AB} (jointly distributed) and A^{BA}, B^{BA} (jointly distributed); no joint distribution across contexts is presupposed. For binary ± 1 variables the within-context product moment equals agreement minus disagreement probability, $E[A^c B^c] = 1 - 2M_c$ for $c \in \{AB, BA\}$, so

$$\big|E[A^{AB}B^{AB}] - E[A^{BA}B^{BA}]\big| = 2\,|M_{BA} - M_{AB}| = 2\,|q_{\mathrm{QQ}}|. \tag{12}$$

Similarly, a ± 1 mean is affine in the "yes" marginal, $E[X^c] = 2p_c(X_y) - 1$, so

$$\big|E[A^{AB}] - E[A^{BA}]\big| = 2\,|OE_A|, \qquad \big|E[B^{AB}] - E[B^{BA}]\big| = 2\,|OE_B|, \tag{13}$$

by the sign conventions of Section 3. The Kujala–Dzhafarov noncontextuality criterion for rank-2 cyclic systems with binary variables [9] states that a maximally noncontextual coupling exists iff the between-context change of the product moment does not exceed the total between-context change of the marginals:

$$\big|E[A^{AB}B^{AB}] - E[A^{BA}B^{BA}]\big| \;\le\; \big|E[A^{AB}] - E[A^{BA}]\big| + \big|E[B^{AB}] - E[B^{BA}]\big|. \tag{14}$$

Substituting the three identities and dividing by 2 yields $|q_{\mathrm{QQ}}| \le |OE_A| + |OE_B| = \mathrm{OSS}$. The stated consequences follow: $q_{\mathrm{QQ}} = 0$ satisfies the inequality for any marginals, and $|q_{\mathrm{QQ}}| > \mathrm{OSS}$ is exactly its failure. □

Proof of Proposition 7. q_{QQ} is a difference of two probabilities, each linear in the corresponding order-conditioned joint; hence q_{QQ} is a linear functional of the pair $J = (\mathrm{joint}_{AB}, \mathrm{joint}_{BA})$. For components $J^{(1)}, \ldots, J^{(K)}$ with $q_{\mathrm{QQ}}(J^{(k)}) = 0$ and weights $w_k \ge 0$, $\sum_k w_k = 1$, applied identically to both coordinates (possibly depending on θ but not on the order),

$$q_{\mathrm{QQ}}\Big(\textstyle\sum_k w_k J^{(k)}\Big) = \sum_k w_k\, q_{\mathrm{QQ}}\big(J^{(k)}\big) = 0. \tag{15}$$

If instead the weights may depend on the order, the mixture value is $\sum_k w_k^{AB} M_{AB}^{(k)} - \sum_k w_k^{BA} M_{BA}^{(k)}$, which need not vanish even when each component satisfies QQ: two components with different, order-invariant disagreement probabilities $M^{(1)} \ne M^{(2)}$, selected with $w^{AB} = (1, 0)$ and $w^{BA} = (0, 1)$, give $q_{\mathrm{QQ}} = M^{(1)} - M^{(2)} \ne 0$. Hence order-dependent selection is a possible, though not necessary, violation channel. □

B Prompts and Item Materials

This appendix documents the prompting conditions, label counterbalancing, and the frozen 18-pair item set used in the pilot. The complete machine-readable materials are provided in `items/pilot_v1_2.json`. The item file was frozen before execution and was not expanded after observing the pilot results.

B.1 System prompts

Track M: direct-evaluation framing. Track M presented the questions with a neutral instruction and no persona framing. For the S1 and S2 items, whose semantic response options were mapped to the neutral labels `A` and `B`, the system prompt (template T-1, reproduced verbatim from the audit record) was:

```
You are asked a series of judgment questions.  For each question, answer with exactly
one letter:  "A" or "B".  Do not explain.
```

For the S3 items, which used the canonical labels `Yes` and `No`:

```
You are asked a series of judgment questions.  For each question, answer with exactly
one word:  "Yes" or "No".  Do not explain.
```

Track P: survey-respondent framing. The single redesign run R1 used the following system-prompt template, with the item-specific reference year substituted before rendering:

```
Answer each question as an average adult American survey respondent in {reference_year}
would have answered.  For each question, answer with exactly one letter:  A or B.
Do not explain.
```

Track P was applied only to the eight year-fixed S1/S2 pairs `s1-01`, `s1-02`, `s1-04`, `s1-06`, `s1-11`, `s1-12`, `s2-01`, and `s2-02`. Their substituted reference years were 1997, 1995, 1996, 1979, 2008, 2012, 1997, and 1995, respectively.

Historical headers in Track M. Two Track M items included an additional historical header. The header was inserted once, before the first question in the sequential interaction, and was not repeated before the second question.

For `s1-05`:

```
The following question is from a survey conducted during the Cold War era.  Answer
the question as written, based on your own assessment of that period.
```

For `s1-11`:

```
The following question is from a survey conducted in 2008.  Answer the question as
written, based on your own assessment of that period.
```

No historical header was added in Track P. In particular, the year-conditioned Track P system prompt replaced rather than supplemented the Track M historical framing.

B.2 Sequential presentation and label counterbalancing

Each pair was presented in both orders, $A \to B$ and $B \to A$. The first response was committed to the conversation history as an assistant message whose content was exactly the accepted canonical label token, before the second question was presented. For each first-answer branch, the history therefore contained the exact canonical response selected for the first question.

For S1 and S2, the two semantic response options were evaluated under two counterbalanced mappings:

$$\begin{aligned} &\text{map-1:} \quad \text{option } 1 \mapsto \texttt{A}, \qquad \text{option } 2 \mapsto \texttt{B}, \\ &\text{map-2:} \quad \text{option } 1 \mapsto \texttt{B}, \qquad \text{option } 2 \mapsto \texttt{A}. \end{aligned}$$

A mapping was held fixed across both presentation orders of a given condition. Before pooling, all token-level probabilities and joint distributions were transformed back to the original semantic option coordinates. The two mappings were then pooled with equal, pre-specified weights. Per-mapping results were retained separately as label-effect diagnostics.

The active label mapping was displayed explicitly to the model in every user turn. Each S1/S2 turn began with a two-line plain-text legend assigning the semantic options to the neutral labels under the active mapping, followed by the question. For example, for item `s1-04` the legend read

```
A = A few
B = Many
```

under map-1, and

```
A = Many
B = A few
```

under map-2. Thus, the model was not required to infer the meaning of the neutral labels from context.

S3 used the canonical labels `Yes` and `No` directly and was analyzed as a separate stratum from the A/B-mapped S1/S2 items.

B.3 Frozen item manifest

Table 5 summarizes the 18 frozen question pairs. A dash indicates that no fixed reference year, matched S1 item, vignette, or historical header was used. The column "P" identifies the eight items included in the Track P redesign run. The full verbatim question wording is contained in the archived JSON file.

Table 5: Manifest of the frozen 18-pair pilot item set. "Sens." indicates the pre-specified sensitive-item flag; "P" indicates inclusion in the Track P redesign run.

ID	Set	Pair type	Semantic options	Year	Sens.	P	Additional material
`s1-01`	S1	consistency	Yes / No	1997	No	Yes	–
`s1-02`	S1	consistency	Yes / No	1995	No	Yes	–

Continued on next page.

Table 5 continued.

ID	Set	Pair type	Semantic options	Year	Sens.	P	Additional material
s1-04	S1	contrast	A few / Many	1996	Yes	Yes	–
s1-05	S1	reciprocity	Yes / No	–	No	No	Cold War header
s1-06	S1	context contrast	Yes / No	1979	Yes	Yes	–
s1-07	S1	reciprocity	Yes / No	–	No	No	–
s1-09	S1	consistency	Approve / Disapprove	–	No	No	–
s1-10	S1	consistency	Approve / Disapprove	–	No	No	–
s1-11	S1	reciprocity	Too personally critical / Not too personally critical	2008	No	Yes	2008 header
s1-12	S1	consistency	Favor / Oppose	2012	Yes	Yes	–
s2-01	S2	consistency	Yes / No	1997	No	Yes	matched to s1-01
s2-02	S2	consistency	Yes / No	1995	No	Yes	matched to s1-02; monitored
s2-09	S2	consistency	Approve / Disapprove	–	No	No	matched to s1-09
s3-01	S3	consistency	Yes / No	–	No	No	vignette
s3-03	S3	contrast	Yes / No	–	No	No	vignette
s3-08	S3	contrast	Yes / No	–	No	No	vignette
s3-09	S3	consistency	Yes / No	–	No	No	vignette
s3-11	S3	mutual obligation	Yes / No	–	No	No	–

B.4 Representative verbatim items

The following examples illustrate the three item strata and the different response formats. Question wording and semantic options are reproduced verbatim from the frozen item file.

Example 1: original survey item with non-Yes/No options (s1-04)

Question A.

Do you think that only a few or many white people dislike black people?

Question B.

Do you think that only a few or many black people dislike white people?

Semantic options: *A few* and *Many*. These options were mapped to A/B under both counterbalanced mappings. The item was marked sensitive and used reference year 1996 in Track P.

Example 2: matched name-substitution item (s2-01)

Question A.

Do you generally think Tony Blair is honest and trustworthy?

Question B.

Do you generally think Gordon Brown is honest and trustworthy?

Semantic options: *Yes* and *No*. The item was matched to `s1-01` and used reference year 1997 in Track P.

Example 3: novel vignette item (`s3-01`)

Vignette.

Two candidates applied for the same software engineering role. Candidate K has strong qualifications but a two-year employment gap. Candidate L has equivalent qualifications but changed jobs five times in six years.

Question A.

Should Candidate K advance to the interview stage?

Question B.

Should Candidate L advance to the interview stage?

The canonical responses for this stratum were `Yes` and `No`; no A/B remapping was applied.

B.5 Complete machine-readable materials

The verbatim wording, semantic response options, metadata, vignettes, historical headers, matching relations, and monitoring flags for all 18 pairs are provided in:

`items/pilot_v1_2.json`

The archived version is identified by the repository commit `4563023`.

The immutable execution records additionally store the rendered system prompt, model and tokenizer revisions, chat-template hash, canonical token identifiers, mapping assignment, random seed, and sampling-check metadata for each reported run.