The Epistemic Triad: Auditing Participant-Endorsed Selections in Hybrid Human–LLM Coding

Iman YeckehZaare[1,2]

[1] Center for Collective Intelligence, Massachusetts Institute of Technology, Cambridge, MA, USA

[2] Honor Education, San Francisco, CA, USA

Author note. This work was conducted as part of the author's Ph.D. dissertation at the School of Information, University of Michigan, Ann Arbor, MI, USA, where the data were collected.

Email: oneman@mit.edu (MIT); iman@honor.education (Honor Education)

Abstract

Hybrid human–LLM qualitative coding is checked against researcher consensus, but participant-centered studies need to know whether automated coding preserves what participants selected about their own responses. We introduce the *Epistemic Triad*, a same-response comparison among participant checklist selections, researcher consensus, and model outputs. Across 879 participants and 1,758 open-text response rows from two passage-comparison questions, eleven confirmatory model comparisons pointed the same way: model outputs were closer to researcher consensus than to participants' selections. On the same responses, seven commercial-API models recovered fewer researcher-marked codes from the full 306-code instrument than from its restricted 61-code subset. On a shared 320-response set, the participant-selected codes one model missed were largely the codes other models missed, exceeding chance with each model's miss-count per response held fixed. Coarser categories weakened but did not erase the participant–model gap. The contribution is an audit of when LLM-assisted coding preserves participant-endorsed checklist selections.

1 Introduction

This paper measures the gap between what participants select when coding their own responses via a structured checklist and what hybrid human–LLM workflows produce. Because models are typically prompted with researcher-designed codebooks, model outputs may sit closer to researcher consensus than to participant self-coding on the same response. The question is how large that gap is, where coverage is lost, and whether the losses are practically consequential (Nelson, 2020; Than et al., 2025; Grimmer and Stewart, 2013).

We term this three-way comparison the *Epistemic Triad*. The participant label is the checklist selection made by the person who wrote the response. The researcher label is the adjudicated researcher consensus. The model label is the set of codes returned by the LLM. Because all three labels are attached to the same response, the audit can ask whether the model is closer to the participant or to the researcher rather than comparing different samples.

The dataset contains $N{=}879$ participants and 1,758 open-text response rows across two parallel passage-comparison questions, labeled Q1 and Q2 in the released data, in which participants selected which passage was easier and then explained their reasoning in un-primed open text. Different analyses use different comparison sets: some need every response with at least one participant selection, some need the responses a given model was able to code, and some need the same responses coded twice under two different code lists. Table 1 names each set, gives its size, and points to where it is used.

We clarify what "participant self-coding" means here. Participants selected codes from the same researcher-designed 306-item checklist used for researcher and model coding, rather than providing open-ended narrative feedback as in member checking (Birt et al., 2016; Carlson, 2010). This design has a tradeoff. Because participants, researchers, and models all use the same checklist, we can compare them on identical responses. Because the checklist is researcher-designed, participant selections are evidence about how participants used this instrument, not a direct measure of everything participants meant.

The closest ancestor is respondent self-coding in open-ended survey analysis (Glazier et al.,

Table 1: Comparison sets used in the main manuscript. Different rows have different counts because the analyses ask different questions: some use every available response, some use only the responses a given model could code, and some require the same responses to be coded under two different code lists.

Comparison set	Responses or response–code pairs counted	Where it is used
Participants only (full corpus)	Responses with at least one active participant selection (n=1,660)	Background on how concentrated participants' code use is across the full corpus (Supplementary Table S16).
Participants vs. researchers vs. each model (per-model matched responses)	Each model's available comparison responses (n=55–337 response rows)	Eleven confirmatory tests of whether each model is closer to researchers or participants (Table 2).
Code concentration (Gini / entropy)	Each model's paired participant–model response set (n=193–775 responses)	Tests of whether each model uses codes more narrowly than participants do.
Shared misses across models	The 320 responses that all nine confirmatory models were able to code	Whether the same participant-selected codes are missed across models.
Same responses under the 61-code list and the full 306-code checklist	Same 337 responses coded twice (once under the 61-code list, once under the full 306-code checklist); 2,204 response–code pairs that researchers agreed should be coded present	Seven-model paired analyses and Figure 1.
Same comparison after collapsing codes into fewer, broader categories	Same comparison after collapsing the 306-code checklist to 20 and 10 categories	Tests whether the pattern (models track researchers more closely than participants) weakens with broader categories.
Newer-model coverage check on the full Q1/Q2 response set (descriptive, not confirmatory)	Full non-empty Q1/Q2 response set (n=1,641) for GPT-5.5 and Gemini 3.1 Pro Preview	Descriptive check of how broadly newer models code, how their selections overlap with the other comparators, and how often they stay within the active code list; separate from the matched-panel confirmatory rows.

2021); our version trades semantic breadth for same-response comparability. Unlike member checking, participants do not endorse researcher interpretations; unlike perspectivist annotation, the three positions are not treated as epistemically interchangeable. We therefore refer to the "participant–model selection gap" rather than the "participant-meaning gap," recognizing that checklist selection is a behavioral proxy that reflects, but does not directly measure, experiential meaning.

The substantive questions are twofold. First, when an LLM is asked to code a qualitative response, does it track the participant's own selections or the researcher's adjudicated schema? Second, when multiple researchers rely on the same small set of models, do their outputs become more concentrated than participant coding and omit the same codes selected by participants? These questions matter because, within the adopted 306-code checklist, participant-centered qualitative designs can treat participant-endorsed selections as validity signals for this checklist-based measure, whereas deployment pipelines are often optimized for standardized, policy-compliant outputs (Ouyang et al., 2022). The latter concern connects hybrid coding to algorithmic monoculture: correlated reliance on the same systems can compress low-prevalence coverage even when top-line performance appears acceptable (Kleinberg and Raghavan, 2021).

2 Related Work

2.1 LLM-Assisted Qualitative Coding and Reflexive Governance

Recent work shows that LLMs can support qualitative coding and related text-classification workflows (Nelson, 2020; Than et al., 2025; Parfenova et al., 2025). At the same time, qualitative methodologists emphasize flexible human coding, adjudication, and reflexive governance rather than automated substitution (Deterding and Waters, 2021; Prahl, 2025). Lei et al. (2025) use LLMs to dynamically cluster qualitative responses within surveys, inviting participants to reflect on emerging themes—the closest existing approach to involving participant voice in real-time, though the design is survey interaction rather than post-hoc same-response audit. Our contribution differs

from these studies by placing participant self-coding, researcher consensus, and model outputs on the same responses; auditing response-level reference provenance; and making confirmatory versus exploratory families explicit in both scripts and release artifacts. This also distinguishes the present study from codebook-following measurement evaluations such as Halterman and Keith (2025), whose five-stage framework asks whether LLMs *follow* a codebook; our estimand is whether the codebook-plus-LLM pipeline *preserves participant-endorsed selections*. Similarly, Tai et al. (2024) test LLM coding consistency across 160 iterations, establishing measurement reliability (LLM-to-LLM and LLM-to-human agreement) but not construct validity against participant self-coding. Our evaluand operates at a different level of the validity hierarchy: whether the coding scheme preserves participant-endorsed selections, a structured proxy for experiential meaning rather than a direct measure of it, not whether the coder applies it reliably. This framing draws on respondent self-coding as a direct audit of semantic validity in open-ended coding (Glazier et al., 2021).

Bias-focused work sharpens why this additional validity layer matters. Ashwin et al. (2025) show that LLM qualitative-analysis errors can be non-random when models are evaluated against researcher-centered labels and interpretations. Our audit targets a complementary layer: whether a codebook-plus-LLM workflow preserves participant-endorsed checklist selections on identical responses, with response-level provenance and aligned comparison sets. It does not estimate generalized LLM bias or causal subgroup harm.

2.2 Label Variation, Perspectivism, and the Participant–Researcher Gap

A growing NLP literature treats annotator disagreement as meaningful signal rather than noise (Plank, 2022; Cabitza et al., 2023; Uma et al., 2021). These frameworks, however, treat all annotators as epistemically interchangeable. The Epistemic Triad introduces *structured asymmetry*: participants have first-person access to the experience that generated the response; researchers designed and adjudicated the checklist; LLMs scale the checklist without experiential grounding. In participant-centered qualitative traditions (Braun and Clarke, 2006, 2019; Lincoln and Guba, 1985), the participant–researcher gap is a construct-validity signal about whether the coding pipeline

preserves participant-endorsed selections.

Recent qualitative-methods work motivates this extension. Schroeder et al. (2025) document researcher concerns about interpretive depth and epistemological compatibility when using LLMs. Hayes (2025) frames LLMs as conversational partners but leaves participants as passive data sources; Dunivin (2025) develops a researcher–LLM–researcher workflow without a participant audit layer. Carlson and Burbano (2026) and Prahl (2025) sharpen governance requirements but neither proposes a *same-response* participant-endorsed-selection preservation audit (i.e., the same participant responses are coded by all comparators in the panel). Across these comparators, validation remains researcher-centric.

2.3 Epistemic Risk and Algorithmic Monoculture

Fricker (2007) conceptualized *hermeneutical injustice* as a structural gap in interpretive resources. We invoke that literature only as a bounded checklist-level risk: hybrid coding may compress low-prevalence participant-endorsed selections within the adopted instrument. Kleinberg and Raghavan (2021) define algorithmic monoculture as correlated reliance on the same systems. We treat monoculture as a motivating hypothesis and estimate observable quantities: distributional concentration, prevalence-band coverage loss, and shared omission of codes selected by participants. As reported below, the evidence is partial—significant for five of eleven confirmatory models (O3, O4-mini, GPT-OSS 120B, GPT-OSS 20B, and Gemini 3.1 Pro Preview)—suggesting model-specific risk rather than a universal property.

2.4 Measurement Error and Inferential Governance

Automated coding introduces measurement error that propagates into downstream claims if provenance, denominator choice, and label-space scope are not audited (Grimmer and Stewart, 2013; Grimmer et al., 2022). Carlson and Burbano (2026) show that classification errors correlated with outcome variables bias downstream inference, reinforcing that correlated LLM errors are an inferential validity problem. Confidence-calibration evidence supports conservative deployment

with abstention when uncertainty is high (Jiang et al., 2020; Zhao et al., 2021). We treat parser fidelity, provenance-clean references, and denominator-scoped estimands as first-class governance requirements. The focus is upstream measurement validity: we audit whether the coding evaluand changes under hybrid workflows rather than asking only whether downstream analyses remain usable.

We therefore frame the novelty claim narrowly and support it with an explicit comparator table rather than prose alone (Supplementary Table S7). Within the comparator set reviewed here, we are not aware of prior work that combines, within one design, the following five components:

1. participant self-coding as the evaluand;
2. same-response participant/researcher/model linkage;
3. response-level provenance auditing;
4. matched-denominator inference; and
5. explicit confirmatory versus exploratory governance.

The nearest alternatives each omit at least one of these components: Ashwin et al. (2025) and Than et al. (2025) evaluate researcher- or benchmark-centered outputs, Halterman and Keith (2025) evaluates codebook adherence, Lei et al. (2025) study participant interaction during data collection rather than post-hoc same-response auditing, and Carlson and Burbano (2026), Prahl (2025), and Schroeder et al. (2025) emphasize governance or researcher practice rather than participant-endorsed-selection preservation measurement (Supplementary Table S7).

3 Methodology

3.1 Data and Epistemic Triad Construction

The data construction proceeds in five stages:

1. Participants (N=879) provided 1,758 Q1/Q2 open-text responses and then self-coded those responses using a 306-code instrument. Of these response rows, 1,660 contain at least one active participant self-selection and define the descriptive participant baseline in Supplementary Table S16.

2. Two researchers independently coded a subset of responses and adjudicated a consensus.

3. For the primary researcher reference, we kept only labels with `in_consensus`=1 and `consensus_human`=1 and complete two-coder adjudication provenance, yielding 838 evaluable responses overall.

4. The eleven-model triad family compares participants, researchers, and models on the responses where that model, the participant selection, and the researcher reference are all available, ranging from n=55 (Gemini 2.5 Pro, limited API availability) to n=337 responses per model (Table 2).

5. Separate response sets support omission, concentration, codebook-expansion, and granularity analyses (Table 1). The paired codebook-expansion family uses the same 337 responses under both code lists, with 2,204 researcher-marked response–code pairs (Figure 1); correlated-omission uses the same n=320 responses across nine high-coverage confirmatory models, and routing calibration remains anchored to the earlier seven-model panel.

Table 1 (and Supplementary Table S3) names each manuscript-facing denominator. Matched-panel GPT-5.5 and Gemini 3.1 Pro Preview rows are included where same-response triad, concentration, shared-omission, and codebook-expansion panels are available. The descriptive GPT-5.2 Thinking result (separate ChatGPT batch route) and the full-Q1/Q2 GPT-5.5/Gemini 3.1 Pro Preview API breadth extension sit outside the confirmatory families and are reported in the Supplementary Material.

We formulate the *Epistemic Triad* by defining three vectors for each response: A (participant self-coding), B (researcher consensus), and C (LLM coding).

3.2 Analytical Strategy

Two analyses follow from the triad geometry: whether the model is closer to the researcher or to the participant, and whether the model's code-frequency distribution is more concentrated than the participant's and whether the codes the model misses overlap across models.

1. **Triad Alignment:** We compute the Jaccard distance between the LLM and the researcher reference with complete two-coder provenance ($D(C, B)$) versus the LLM and the participant ($D(C, A)$), where a smaller distance indicates higher alignment. For the confirmatory family (the eleven model configurations on which the hypothesis tests are run, with Benjamini–Hochberg multiplicity correction across the family), we test $D(C, B) < D(C, A)$ using Wilcoxon signed-rank tests with Benjamini–Hochberg correction across eleven comparisons, reporting Cohen's d_z and bootstrap 95% CIs. If $D(C, B) < D(C, A)$, the model output is closer to researcher consensus than to participant self-coding on this evaluand.

2. **Distributional Concentration and Correlated Omission:** We compare code-frequency distributions using Gini coefficient and Shannon entropy, with paired deltas on each model's available comparison responses and Benjamini–Hochberg correction across eleven confirmatory model comparisons. To localize omission risk, we audit pairwise overlap in missed codes selected by participants on shared confirmatory intersections against independence and against a shuffle that keeps each model's number of misses per response fixed but changes which participant-selected codes are missed (the response-preserving null), so excess overlap above the null is correlated omission rather than per-response capacity.

Full statistical specifications (bootstrap resamples, permutation counts, seed values), the seven-diagnostic audit framework, and the denominator crosswalk are in the Supplementary Material (Section 1; Tables S5, S3).

4 Results

Not all results in this section carry the same evidentiary weight. The pre-specified statistical tests are the eleven participant–researcher–model alignment comparisons and the eleven model-specific concentration tests, with Benjamini–Hochberg correction (Supplementary Material). Two newer models, GPT-5.5 and Gemini 3.1 Pro Preview, enter those confirmatory families only on the responses where every comparator in the panel (participants, researchers, and the model under test) coded the same items — the matched-panel condition — because that is the only subset on which the relational, concentration, and codebook-expansion comparisons can be computed without changing the denominator. Two other results are descriptive rather than confirmatory and are reported in the Supplementary Material: the GPT-5.2 Thinking run (separate ChatGPT batch collection route) and the full-Q1/Q2 GPT-5.5 and Gemini 3.1 Pro Preview breadth extension (different response set). Semantic audit packets, which compare each model's free-text justification with researcher and participant intent, remain background calibration only, because their labels have not yet been independently dual-coded. The matched-panel frontier semantic gate (which requires independently dual-coded participant-selected, model-only, and overlap codes) is not yet passed for GPT-5.5 or Gemini 3.1 Pro Preview because that dual coding is not yet available. We therefore do not claim causal subgroup harm, route-neutral model ranking, semantic recovery by frontier models, or cross-domain portability from this single-case demonstration.

4.1 RQ1: Triad Alignment

Our first hypothesis posited that LLMs more closely approximate researcher schemas than participant selections. This subsection concerns the eleven main comparisons, each using researcher labels with complete two-coder adjudication provenance and the responses available for that model. In plain terms, every main model comparison shows the same pattern: model outputs are closer to researcher consensus than to participant self-coding. Because Jaccard distance is lower when two code sets overlap more, the consistently smaller $D(C, B)$ values mean that model outputs track

researchers more closely than participants.

Table 2 reports the paired tests. The Jaccard distance between the LLM and the researcher ($D(C, B)$) was significantly *smaller* than the distance between the LLM and the participant ($D(C, A)$); all $q_{\mathrm{BH,conf}} < .001$ after Benjamini–Hochberg correction across the confirmatory family. Supplementary Table S4 adds the direct participant–researcher benchmark on those same model-specific response sets: mean $D(A, B)$ spans 0.843–0.873 across confirmatory model comparisons, clarifying that the models converge toward a researcher reference that is itself well separated from participant self-coding.

Across confirmatory model comparisons, effect sizes range from small ($d_z = -0.293$ for GPT-OSS 20B) to large ($d_z = -0.975$ for GPT-5), with a median $|d_z| = 0.770$. For example, Gemini 3 Flash showed $\bar{D}_R = 0.706$ versus $\bar{D}_P = 0.913$ ($\Delta = -0.207$, $d_z = -0.942$). The pattern is consistent across all three closed-model vendors and both open-weight models. Supplementary Table S14 confirms directional consistency under the archival mixed-reference view.

4.2 RQ2: Distributional Concentration and Correlated Omission

Our second research question maps three distributional questions that use different denominators. Concentration uses model-specific matched response sets. Shared omission uses the common n=320 response panel. Codebook expansion uses the same 337 responses under two code lists.

Concentration: do models narrow the distribution of selected codes? The concentration analysis asks whether model outputs use selected codes less evenly than participant selections. The unit is each model's matched response set, and the counted codes are the codes selected by participants for those responses. Five of eleven models — O3, O4-mini, GPT-OSS 120B, GPT-OSS 20B, and Gemini 3.1 Pro Preview — show significant compression on both Gini and entropy after Benjamini–Hochberg correction; the remaining six do not. This is partial, model-specific evidence rather than a universal monoculture claim. Among the compressing models, ΔGini ranges from $+0.014$ to $+0.043$ and ΔH from -0.594 to -0.172 (all $q \leq .007$); per-model values, shared-

Table 2: Epistemic Triad: mean Jaccard distance from each LLM to researcher consensus with complete two-coder provenance ($\bar{D}_R$) and to participant self-coding ($\bar{D}_P$). Lower distance = stronger alignment. For each model comparison, n is the number of responses with participant self-coding, adjudicated researcher consensus with complete provenance, and model predictions. All eleven confirmatory model comparisons are significantly closer to researcher consensus than to participant self-coding ($\bar{D}_R < \bar{D}_P$); the exploratory GPT-5.2 Thinking row follows the same direction. $\Delta = \bar{D}_R - \bar{D}_P$ (negative = closer to researcher consensus). $q_{\mathrm{BH,conf}}$: Wilcoxon signed-rank with Benjamini–Hochberg correction across the eleven confirmatory model comparisons only. 95% CIs: 2,000 bootstrap resamples. Rows are conservative comparisons with harmonized denominators; GPT-5.2 Thinking (†) is reported as descriptive sensitivity evidence (no inferential test; legacy mixed-reference sensitivity results and confirmatory matched-condition inferential families are reported separately in the Supplement).

Model	n	$\bar{D}_R$ [95% CI]	$\bar{D}_P$ [95% CI]	Δ	$q_{\mathrm{BH,conf}}$	d_z
O3	325	0.920 [0.898, 0.941]	0.985 [0.977, 0.991]	−0.0652	< .001***	−0.351
O4-mini	335	0.905 [0.888, 0.921]	0.967 [0.960, 0.973]	−0.0618	< .001***	−0.400
GPT-5-mini	336	0.730 [0.706, 0.752]	0.909 [0.896, 0.921]	−0.1785	< .001***	−0.770
Claude Sonnet 4.5	333	0.717 [0.695, 0.740]	0.912 [0.898, 0.925]	−0.1953	< .001***	−0.883
Gemini 3 Flash	337	0.706 [0.681, 0.729]	0.913 [0.901, 0.924]	−0.2075	< .001***	−0.942
GPT-5	132	0.703 [0.663, 0.740]	0.923 [0.904, 0.940]	−0.2196	< .001***	−0.975
Gemini 2.5 Pro	55	0.755 [0.692, 0.813]	0.935 [0.912, 0.958]	−0.1802	< .001***	−0.801
GPT-OSS 120B	332	0.779 [0.752, 0.805]	0.932 [0.918, 0.944]	−0.1526	< .001***	−0.660
GPT-OSS 20B	320	0.932 [0.911, 0.951]	0.980 [0.970, 0.988]	−0.0474	< .001***	−0.293
GPT-5.5	334	0.718 [0.693, 0.742]	0.913 [0.901, 0.925]	−0.1945	< .001***	−0.855
Gemini 3.1 Pro Preview	321	0.901 [0.876, 0.924]	0.975 [0.964, 0.983]	−0.0733	< .001***	−0.378
GPT-5.2 Thinking†	337	0.838 [0.816, 0.858]	0.945 [0.933, 0.955]	−0.1063	—	− − −

intersection sensitivity, prevalence-band diagnostics, and a compact confirmatory-core synthesis are in the Supplementary Material (Tables S18–S17, S24; Figure S2).

Shared omission: do models miss the same participant-selected codes? The shared-omission analysis asks whether models miss the same participant-selected codes, not merely whether each model misses many codes. We evaluate this on the same n=320 responses across nine high-coverage confirmatory models, with missed items counted as response–code pairs selected by participants. The models' missed-code sets overlap strongly (mean pairwise Jaccard = 0.932 [0.927, 0.938]), meaning that when one model misses a participant-selected code on a given response, the other models tend to miss the same code on the same response. To check that this overlap is not just a side-effect of each model missing many codes, we compare the observed overlap against a shuffle that holds each model's number of misses per response fixed but randomizes which participant-

selected codes are recorded as missed (the response-preserving null). Observed overlap exceeds the null by $+0.038$ [0.033, 0.042]. This supports a bounded claim that the nine models miss the same participant-selected codes more often than chance on the common response set; it does not estimate a general missed-meaning rate (Supplementary Table S19).

Codebook expansion: what changes when models see the full checklist? The codebook-expansion analysis asks what happens when the same responses are coded against the restricted 61-code list and the full 306-code checklist. The unit is the common n=337 response subset coded under both code lists, which contains 2,204 response–code pairs that researcher consensus marked present. We use *recall* in its standard same-response form: the share of those 2,204 researcher-marked pairs that a model also marked present on the same response. All seven commercial-API models recover a smaller share of those pairs when the active code list expands from 61 to 306 codes, with the largest loss for O4-mini ($\Delta R = -0.089$ [-0.106, -0.074]; Figure 1). GPT-5.5 also shows a significant loss ($\Delta R = -0.026$ [-0.037, -0.015]), while Gemini 3.1 Pro Preview is near zero and statistically indistinguishable from no change ($\Delta R = -0.004$ [-0.015, +0.007]). The subset is more participant-aligned than the broader pool (participant conservative recall 0.326 vs. 0.108), but post-stratified recall-delta shifts remain $\leq$0.005, so this remains a same-response stress test rather than a full-pool estimate.

4.3 Generalizability to Coarser Codebook Sizes

The granularity analysis asks a simple reader question: is the main result only caused by the very detailed 306-code checklist? The answer is no. The model–researcher direction remains at coarser levels, but it weakens and the concentration result no longer generalizes cleanly.

To test this, we repeated the triad alignment and concentration analyses at theme-level granularity (20 categories) and minimal-level granularity (10 categories) for the full eleven-model confirmatory family plus one descriptive GPT-5.2 result (Supplementary Tables S26–S27). At theme-level granularity, all eleven confirmatory models remain closer to researcher consensus than to participant

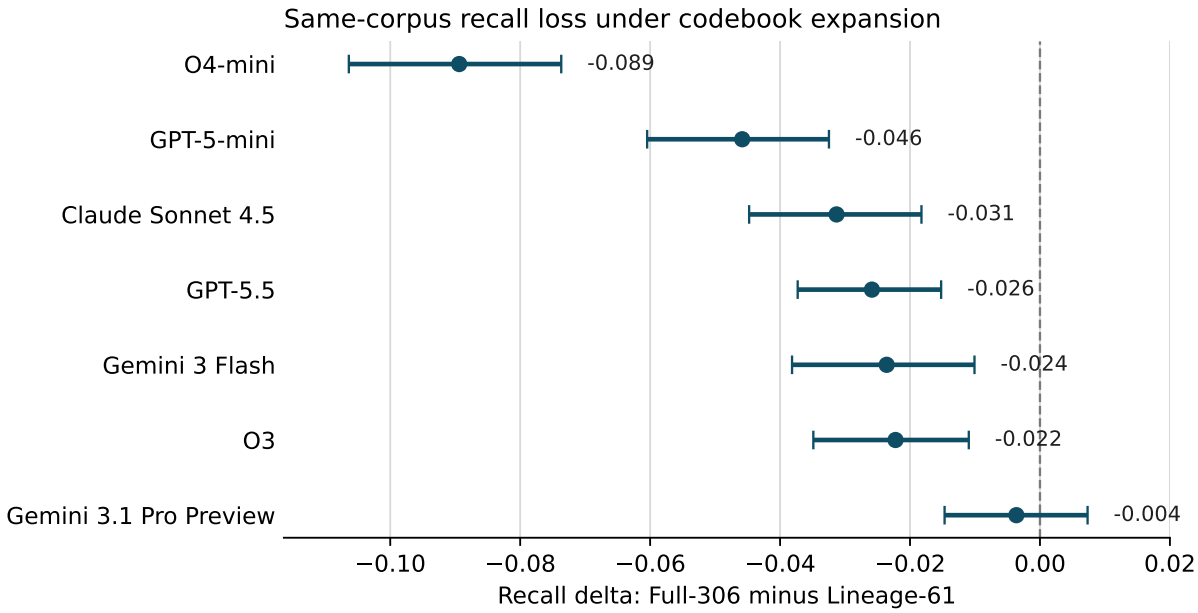

Figure 1: Same-corpus recall deltas under codebook expansion on the confirmatory n=337 responses. Points show the recall difference (full 306-code instrument minus restricted 61-code subset) for each commercial-API model, with bootstrap 95% CIs. The common subset is coded under both code lists and is more participant-aligned than the broader n=838 evaluation pool (participant conservative recall 0.326 vs. 0.108); post-stratified sensitivity analyses across the original five matched-condition models preserve sign and shift recall deltas by only 0.001 to 0.005. Across the seven API models, every point estimate is negative once active code exposure expands from 61 to 306 codes; the largest loss is for O4-mini ($\Delta R = -0.089$ [-0.106, -0.074]), while Gemini 3.1 Pro Preview is near zero and its interval crosses zero.

self-coding after Benjamini–Hochberg correction within the theme-level family ($q_{\mathrm{BH}} < .001$ for all eleven), with effect sizes ranging from small (GPT-OSS 20B: $d_z = -0.271$) to large (Gemini 3 Flash: $d_z = -0.791$). At minimal-level granularity (10 categories), the same direction remains for all eleven confirmatory models, but only seven rows retain within-family significance after correction (all $q_{\mathrm{BH}} \leq .001$); the remaining four are directionally negative but indistinguishable from zero. The GPT-5.2 row is reported descriptively only in the Supplementary Material.

The pattern weakens at coarser codebook sizes, but it does not reverse. The 306-code concentration pattern does not replicate cleanly at coarser levels: at theme-level granularity, Gini shifts are mixed and entropy increases in nine of eleven confirmatory model comparisons; at minimal-level granularity, both metrics remain heterogeneous. A descriptive intermediate collapse ladder at approximately 193, 129, 74, 20, and 10 categories shows smooth recall gains for participants and the three high-coverage anchor models (O3, O4-mini, Gemini), with per-step values in Supplementary Table S28. The separate minimal 10-category single-label ablation reaches 79.7–86.0% model recall, isolating multi-label disambiguation cost rather than providing a direct estimate for theme-scale workflows (Supplementary Table S29). Supplementary Table S30 localizes the same weakening at the instance level: across three high-coverage anchor models (Gemini, O3, O4-mini), participant and model true-positive selections never coincide on a consensus-positive response–code pair above 10 categories; overlap appears only after collapse to 10 categories.

A theme-level routing simulation tests whether routing signals retain value at 20 categories (Supplementary Tables S31–S32). Six of seven anchors fall within the matched random-baseline 95% interval; only O4-mini sits modestly above it by 2.0 percentage points. At theme-level granularity, routing is therefore a workload-bound diagnostic rather than a validated gain-over-random signal.

5 Discussion

The practical lesson is straightforward. If a project only needs to reproduce researcher consensus, LLM coding may look acceptable. If a project needs to preserve participant-endorsed checklist selections, researcher-consensus reproduction is not enough.

Three evidence pillars support that lesson. First, all eleven confirmatory models are closer to researcher consensus than to participant self-coding. Second, five models use selected codes in a more concentrated way than participants do. Third, on the common response set the participant-selected codes one model misses are largely the codes other models miss, and this overlap exceeds the shuffle that keeps each model's number of misses per response fixed. Together, these results show that model performance depends on which epistemic position the workflow is supposed to preserve.

The contribution is not a model ranking. It is an audit design that identifies where participant-endorsed selections drop out and which denominator supports each claim. The participant–researcher baseline on the same comparison responses matters because models converge toward a researcher reference that is itself far from participant self-coding. Reproducing researcher consensus is therefore not equivalent to preserving participant selections.

The boundaries are equally important. Participant self-coding is a checklist-based proxy, not a direct measure of all participant meaning. The manuscript is a single-case methods demonstration, not a cross-domain validation of framework portability. Semantic packets are calibration-only unless independently validated; the independently dual-coded 35-row remap audit validates parser fidelity only. The exploratory GPT-5.2 extension (Supplementary Tables S35–S37) shows heterogeneity tied to collection route rather than a confirmatory revision. On the shared $n{=}320$ panel, GPT-5.2 added activity is burden-dominant and is used for provenance, burden localization, and bounded remap validation only. The matched-panel GPT-5.5 and Gemini 3.1 Pro Preview rows support relational alignment, concentration, and codebook-expansion estimands only; their semantic gate did not pass for recovery or safety claims. The full-Q1/Q2 frontier API breadth extension (Supplementary Tables S40–S42) remains descriptive: GPT-5.5 and Gemini 3.1 Pro Preview both recover fewer

researcher-present pairs with complete two-coder provenance than participant+model union coding recovers, and the Gemini run required substantial filtering to active codes before analysis.

For codebook-based projects, researcher-consensus reproduction may suffice when the estimand is researcher consistency; for participant-centered projects, this manuscript supports a quantitative warning about correlated omission, not a validated missed-meaning rate. Routing rules should be treated as workload-bound screens; at theme-level granularity, six of seven anchors remain inside the matched random-review interval.

Conclusion. The Epistemic Triad shows that, on this checklist-based measure, hybrid human–LLM coding can reproduce researcher categories more readily than participant-endorsed selections. Its value is as an audit framework: it makes denominator choice, concentration risk, omission overlap, and semantic calibration explicit before a research team treats automated coding as methodologically acceptable. The confirmatory quantitative layer is the basis for that conclusion; the same-analyst and single-pass semantic packets remain calibration-only layers rather than validation layers, and the paper remains a single-case methods demonstration rather than a portability validation.

6 Limitations

Five limitations bound the scope of the present work:

1. The evaluation covers a single educational domain. Generalization to clinical, ethnographic, or organizational contexts requires replication with participant self-coding in those settings; Supplementary Table S34 reruns the audit machinery on a public task but lacks participant self-coding.

2. Confirmatory denominators are fractions of the full participant pool ($N{=}879$); per-analysis denominators are listed in Table 1. The codebook-expansion subset is more participant-aligned than the broader pool (conservative recall 0.326 vs. 0.108); post-stratified sensitivity

analyses preserve the sign of the original five model-recall deltas and shift them by at most 0.005. Broader transport guardrails for other metrics are reported in the supplement.

3. The study was not pre-registered; the Epistemic Triad framework was developed iteratively alongside data collection.

4. Semantic validation is uneven. The shared-omission and GPT-5.2 co-selection packets lack independent re-auditing, whereas only the 35-row remap audit is independently dual-coded. The GPT-5.5/Gemini 3.1 Pro Preview matched-panel semantic gate did not pass; those rows support relational, concentration, and codebook-expansion claims only, while the full-Q1/Q2 API breadth extension is parsed and filtered to active codes but not independently semantically re-audited.

5. The gap between structured checklist selection and open-ended respondent validation (Birt et al., 2016) remains a design constraint: participant-endorsed selections are a behavioral proxy for experiential meaning, not a direct measure of it.

Author Contributions

The sole author conceptualized the study, designed and conducted the analyses, curated the release artifacts, drafted and revised the manuscript, and approved the final manuscript.

7 Statements and Declarations

Ethical considerations

This study was reviewed by the University of Michigan Institutional Review Board Health Sciences and Behavioral Sciences (IRB-HSBS) under protocol HUM00201762, “How do Linear vs. Non-linear vs. hybrid representations of text affect scientific reading comprehension?” (amendments Ame00117430 and Ame00122854). The IRB-HSBS issued the verbatim determination: “As part of

their review, the University of Michigan Institutional Review Board Health Sciences and Behavioral Sciences has determined that this study is no more than minimal risk and exempt from ongoing IRB oversight." The determination corresponds to Exemption Category 3 under 45 CFR 46.104(d)(3) (benign behavioral interventions; verbal or written responses; no physiological data collection; prospective agreement to participate). The exempt determination was effective August 27, 2021. All study procedures complied with the determination's conditions, including the 45 CFR 46.104(d)(3) requirements for benign behavioral interventions and prospective participant agreement.

Consent to participate

All participants provided electronic prospective agreement before participation, via the online informed-consent form rendered on the experiment website (`https://visualexp1.web.app/`; copy of record retained in the IRB-HSBS protocol packet for HUM00201762). The consent modality (prospective agreement without a signed document) was authorized under the IRB-HSBS exempt determination at 45 CFR 46.104(d)(3) for benign behavioral interventions. The verbatim institutional risk, voluntariness, and data-sharing sentences presented to participants were: "There are no risks, however, you will be asked to share your screen to present how you read the passages and answer questions, and talk throughout the recorded interview." "Participating in this study is completely voluntary. Even if you decide to participate now, you may change your mind and stop at any time. You may choose not to answer any or request to pause or stop the interview for any reason." "We will not collect any sensitive or identifiable information." "Information collected in this project may be shared with other researchers, but we will not share any information that could identify you."

Consent for publication

Not applicable. This manuscript does not publish raw participant open text, direct identifiers, or individual-identifying details. Open-text responses are stored under institutional data governance controls; participant identifiers have been replaced with pseudonymized labels throughout released

materials.

Declaration of conflicting interest

The author declared no potential conflicts of interest.

Funding statement

No external funding was received for this research, authorship, or publication.

Data availability

Analysis scripts, aggregated CSV artifacts, prompt templates, audit labels, and codebook snapshots accompany this manuscript in the released artifact package. The supplement documents the artifact entry points, claim-to-artifact traceability index, and claim-validation audit. Raw participant open text and direct identifiers remain restricted under institutional data governance. LLM prompts, model identifiers, and inference parameters are documented in Supplementary Section 1.

Generative AI use

Generative AI systems were used as study objects in the reported evaluations and for limited drafting and editing support under author supervision. All numerical claims were checked against regenerated claim-validation artifacts (all matched), and cited-reference verification is reported record-by-record in Supplementary Section 6. The author remains responsible for the final content.

References

Ashwin, J., Chhabra, A., and Rao, V. (2025). Using large language models for qualitative analysis can introduce serious bias. *Sociological Methods & Research*, 54.

Birt, L., Scott, S., Cavers, D., Campbell, C., and Walter, F. (2016). Member checking. *Qualitative Health Research*, 26(13):1802–1811.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Braun, V. and Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4):589–597.

Cabitza, F., Campagner, A., and Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 6860–6868. AAAI Press.

Carlson, J. A. (2010). Avoiding traps in member checking. *The Qualitative Report*, 15(5):1102–1113.

Carlson, N. A. and Burbano, V. (2026). The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*.

Deterding, N. M. and Waters, M. C. (2021). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological Methods & Research*, 50(2):708–739.

Dunivin, Z. O. (2025). Scaling hermeneutics: a guide to qualitative coding with LLMs for reflexive content analysis. *EPJ Data Science*, 14:28.

Fricker, M. (2007). *Epistemic Injustice*. Oxford University Press.

Glazier, R. A., Boydstun, A. E., and Feezell, J. T. (2021). Self-coding: A method to assess semantic validity and bias when coding open-ended responses. *Research & Politics*, 8(3):1–8.

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, Princeton, NJ.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

Halterman, A. and Keith, K. A. (2025). Codebook LLMs: Evaluating LLMs as measurement tools for political science concepts. *Political Analysis*, pages 1–17. FirstView article.

Hayes, A. S. (2025). "conversing" with qualitative data: Enhancing qualitative research through large language models (LLMs). *International Journal of Qualitative Methods*, 24.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Kleinberg, J. and Raghavan, M. (2021). Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118.

Lei, K., Ladenburg, A., Petiwala, Z. K., Wang, Z., Jhawar, D., Bisht, I., Kumar, A., and Lee, D. T. (2025). Dynamic surveys: Using LLMs to blend qualitative depth, quantitative structure, and collaborative interaction. *Proceedings of the ACM on Human-Computer Interaction*, 9(CSCW).

Lincoln, Y. S. and Guba, E. G. (1985). *Naturalistic Inquiry*. Sage Publications, Newbury Park, CA.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1):3–42.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Parfenova, A., Marfurt, A., Pfeffer, J., and Denzler, A. (2025). Text annotation via inductive coding: Comparing human experts to LLMs in qualitative data analysis. In *Findings of the Association*

for Computational Linguistics: NAACL 2025, pages 6471–6484, Albuquerque, NM. Association for Computational Linguistics.

Plank, B. (2022). The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

Prahl, A. (2025). The AI-reflexivity checklist (ARC): A pre-analysis pause for LLM-assisted coding. *Qualitative Health Research*. Online ahead of print.

Schroeder, H., Aubin Le Quéré, M., Randazzo, C., Mimno, D., and Schoenebeck, S. (2025). Large language models in qualitative research: Uses, tensions, and intentions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., and Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, 23:1–14.

Than, N., Fan, L., Law, T., Nelson, L. K., and McCall, L. (2025). Updating "the future of coding": Qualitative coding with generative large language models. *Sociological Methods & Research*, 54(3):849–888.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.