Encevra Research Round 45: The Rule or the Reader? Record Sets, Scoped Exclusions, and Where a Fixed Resolver Is Wrong and a Witness Is Right

Date: 2026-11-03
Research lab: Encevra Research
Author: Ben Wilkins, Encevra (ORCID 0009-0005-6532-3200)
Short code: `RSE-45` (Record Sets and Scoped Exclusions)
Primary tracks: Reasoning and Verification ("How does a system know when an answer is ready to release?"), Collaborative track ("When does dividing reasoning across specialists help, and when does it fragment evidence?")
Position in the program: fifth round of the Round 41 to 50 "Theory v1" block. The roadmap item for Round 45 is resolver failure: where fixed rules are wrong and witnesses are right (proposition P2).
Carry-forward boundary: Round 44 (`RLS-44`) found that a record level ledger, read by deterministic code, held near miss items that condition ledgers and direct readers released as unsafe FAILs, and named two weak points of its own resolver: requirements met only by two records read together, and negative statements inside one record read as statements about every record. Its handoff asked whether a ledger with record sets and a scoped exclusion field fixes both, where the fixed rules are wrong while a witness reading the records together is right, and whether a completeness record, rather than any added record, drives the FAIL on completeness items. Round 45 builds that ledger, compares four fixed rule sets on identical witness output, adds a control variant for the completeness question, and fixes an exhibit track on which a rewriting validator can be wrong. It does not rerun the Round 44 record level study, the Round 43 convention study, or the Round 42 role pair design.

Abstract

A deterministic resolver is only as right as its rules. Round 44 showed that reading evidence ledgers with fixed rules stops readers from treating a defective record as proof that no valid record exists, and it also showed where those rules were too narrow: a requirement met by two records together could never be MET, and any quoted negative statement made a condition UNMET, whatever it covered. Round 45 asks the roadmap's question directly. When the reader gets the records right, where is the fixed rule wrong, and when the rule is changed to fix that, what new mistakes does it let through?

We wrote six new contract families (an escrow release needing joint instructions, a preferred stock protective provision needing a named director's written consent, a termination notice to two addressees, payment and performance bonds of stated sums, an environmental report with a reliance letter, and tenant estoppels covering 75 percent of rentable area), each in five variants: K2 (two records together meet the requirement and neither does alone), T (a near miss record), TC (the near miss plus a completeness record), TX (the near miss plus an added record that is not a completeness record), and TS (the near miss plus an exclusion whose scope leaves a qualifying route open). Every label carries a two completion certificate and all 30 packets were reviewed blind under an open and a closed record before any test call. The new ledger asks each witness for a status per record, for record sets (one

quote from each record that together meet the requirement), and for any exclusion with who states it, what it rules out, the period it covers, and whether it covers the requirement. The same ledger is then resolved by four rule sets: the Round 44 rules (R44), R44 with record sets (R44+S), R44 with the scope test (R44+X), and both (R45, the pre registered primary rule), plus one exploratory rule declared before any call (R45+L).

We report three strictly separated evidence categories.

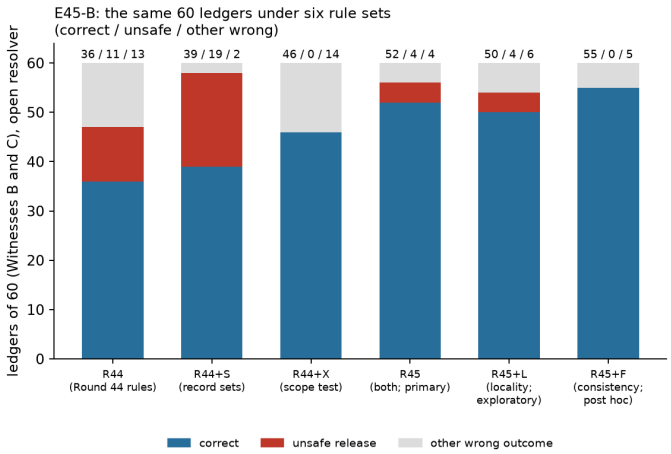
Deterministic (`E45-A`, no model calls). All six family certificates pass. The variants differ from T only by the added record; exactly the T, TX, and TS items are HOLD under an open record. On gold ledgers, R45 reproduces all 60 labels under both resolvers; R44 is wrong on all 6 K2 items under both and releases an unsafe FAIL on all 6 TS items under the open resolver. On ledgers carrying a single scripted reader mistake, a record scoped negative quoted as an exclusion with covers_requirement false is caught by the scope test, but the same quote with covers_requirement true defeats every pre registered rule, and only the exploratory locality rule catches it. On the 24 new SEC exhibits, a rewrite that follows a misleading reviewer note produces an unsafe release on all 12 HOLD exhibits, which the unchanged rewrite gate routes to review, and a false hold on all 12 CERTIFY exhibits, which the gate accepts.

Live authored (`E45-B`, 6 families, 30 items, 152 completed calls). The blind review agreed with all 60 authored labels and reported 7 defect notes, one or more in every family, so under the fixed adjudication rule every family is marked contested. Witnesses B and C returned 60 record set ledgers. Resolved by R45 with the open resolver, 52 of 60 were correct with 4 unsafe releases (Witness C 30 of 30 correct; Witness B 22 of 30 with 4 unsafe). The same 60 ledgers under the Round 44 rules gave 36 correct and 11 unsafe. On 19 ledgers R44 was wrong and R45 right: 9 K2 ledgers where the witness listed a valid record set, and 10 ledgers where a negative statement the witness itself marked as not covering the requirement no longer produced FAIL. On 3 ledgers R44 was right and R45 wrong, all from Witness B. All 4 R45 unsafe releases are Witness B on the tenant estoppel family, the round's most instructive failure: the witness correctly marked a stale certificate FAILS under the date limit, then listed the same certificate in a record set for the 75 percent threshold, and R45 released PASS. The completeness control answered Round 44's second question: under R45, TC items resolved to FAIL on 8 of 12 ledgers and TX, TS, and T items on 0 of 12 each. The same two witnesses reading the packets directly under an open declaration were correct on 46 of 60 with 14 unsafe FAILs, 9 of them on plain T near miss items. Where the R45 resolver was wrong (8 ledgers) the direct reader was right on 7; where the direct reader was wrong (14) the resolver was right on 13. Witness D, as a separate panel, was correct on 22 of 30 under R45 with no unsafe release and held all 6 TC items.

Real exhibits (`E45-C`, 72 calls). Witness C rewrote 24 seeded ledgers and 24 annotated ledgers carrying a misleading reviewer note, and ran a flag only check on the annotated ledgers. All 72 outcomes were correct. The rewriting validator repaired every seeded ledger and followed the misleading note on none of the 24 annotated ledgers; the flag only check flagged 10 of the 12 HOLD annotated ledgers and none of the CERTIFY ones. The unchanged gate routed the 12 correct CERTIFY repairs to review, its only cost on this track.

Completion note. The run budget rule first stopped the round after 57 of 60 record set calls. The 3 missing calls, the direct reader phase, the exhibit validator phase, Witness D, and the synthesis review were rerun the same day, after the first results had been read, with unchanged prompts, settings, and phase order (Section 3.5). The rerun is disclosed as post hoc; no completed call was repeated and no call was selected on outcomes.

Interpretation, kept separate. The fixed rules were wrong in exactly the two places Round 44 named, and adding record sets and a scope test fixed both on these families for a strong witness. The rules were also wrong in a third place nobody had named: the resolver treats each condition a witness lists as independent and calls a condition MET if any record satisfies it, so a requirement split into "each certificate dated within 30 days" and "certificates covering 75 percent" can be met piecewise by different records. In that case the witness's record statuses were right and the rule's existential reading was wrong. A post hoc consistency test (a record set may not use a record the same ledger marks FAILS) removes all 4 unsafe releases on these ledgers; it is reported only as a hypothesis for a later round. Adding one rule feature without the other made things worse: record sets alone raised unsafe releases from 11 to 19. Resolver and direct reader errors barely overlapped (1 ledger wrong under both), which suggests the two are complementary checks rather than substitutes: direct readers still read a near miss as FAIL, and the resolver fails where its rules cannot express the requirement.



Same ledger, two fixed rule sets (open resolver):
R44 wrong and R45 right on 19; R44 right and R45 wrong on 3

	R45 right	R45 wrong
R44 right	33	3
R44 wrong	19	5

All 3 reverse cases are Witness B. In 2 (tenant estoppels) a record set counted a certificate the same ledger marked FAILS; in 1 the witness judged a completeness record out of scope. Direct readers on the same items: 46 of 60 correct, 14 unsafe; resolver wrong and reader right on 7, resolver right and reader wrong on 13.

---

1. Archive continuity and anti-repetition verification

Before designing any Round 45 experiment we rebuilt the archive ledger from the live research registry (`prior-rounds-ledger.json`, `38` indexed rounds, `179` registry hypotheses, `70` registry experiments). A keyword scan of titles, questions, summaries, findings, open questions, hypotheses, and experiments for the Round 45 constructs returned:

- "record set" and "scoped exclusion" only in Round 44's recommendation and open question, which this round carries out;
- "fixed rule", "resolver failure", and "together satisfy" only in Round 44's summary and open question;
- "joint" in Rounds 14, 15, 26, 33, 41, and 42, about joint misreading across roles and joint decoding settings; none concerns two records that jointly meet one requirement;
- "direct reader" in Rounds 42 to 44; the open declaration prompt is reused unchanged and only as a baseline;

- "counterpart" in Rounds 38 and 41; here a counterpart appears as a near miss record in an escrow release (delivered by a sender not on the authorized list), a new setting;
- "wrong rewrite" only as a Round 44 hypothesis, where none occurred.

"Rule or the reader", "annotated ledger", "reviewer note", "misleading note", "escrow", "written consent", "reliance letter", "Phase I", "estoppel", and "payment bond" returned no hits. The full table is in `anti-repetition-verification.md`.

Prior round	Focus	How Round 45 differs
`RLS-44` (Round 44)	Record level status; completeness records; blind label review; gated rewrite on seeded ledgers	Adds record sets and a scoped exclusion field; resolves identical ledgers under four rule sets; adds TX and TS controls; adds annotated exhibit ledgers on which a rewrite can be wrong
`RCV-43` (Round 43)	Declared open and closed record conventions	Conventions used only to define labels
`RDX-42` (Round 42)	Role against witness decorrelation	Not rerun
`NCT-40` (Round 40)	Six topologies; the roadmap	Round 45 is the roadmap item on resolver failure

All 24 exhibits come from new EDGAR full text searches and overlap none of the 1,169 accessions or 902 filer names used in Rounds 39 to 44.

2. Where this round sits in the program

Round 40 proposed P1 to P5. P2 states that release decisions should be made by a deterministic resolver over evidence that witnesses extract, rather than by witnesses voting on verdicts. Rounds 41 to 44 tested what witnesses extract; Round 44 tested a resolver over record level statuses and found its two narrow rules. The roadmap item for Round 45 tests P2 at its weakest point: the resolver's rules are fixed in advance, and a fixed rule can be wrong where a careful reader is right. If the resolver fails more often than readers do when they read the same records, P2 needs revising; if the failures are few and nameable, P2 survives with named exceptions.

3. Design

3.1 Families and variants

Family	Requirement	Near miss record in T	Scope gap in TS
`M01` escrow release	joint written instruction signed by Buyer and Seller, delivered by an authorized sender	Seller's counterpart delivered by a person not listed on Schedule B	the agent's certificate rules out further Buyer counterparts only

Family	Requirement	Near miss record in T	Scope gap in TS
`M02` protective provision	board approval including the written consent of the Series A Director	the director's email stating it is not her written consent	the minutes rule out approvals at board meetings only, not written consents
`M03` termination notice	notice delivered to the CEO and to the General Counsel	General Counsel delivery returned undelivered	the courier's statement covers its own deliveries only; hand delivery is allowed
`M04` bonds	payment and performance bonds each in the stated penal sum	payment bond at half the stated sum, no rider attached	the register covers bonds from one surety only
`M05` environmental report	Phase I report plus a reliance letter naming the lender	letter naming the borrower only	the consultant's statement covers letters issued before an earlier date only
`M06` tenant estoppels	estoppels from tenants leasing at least 75 percent, each dated within 30 days of closing	a certificate dated 51 days before closing, no bring down attached	the seller's statement covers certificates through a date ten days before closing

Each family has five variants. In K2, the requirement is met by two records together and neither meets it alone (for example, one tenant's 45 percent estoppel plus another's 30 percent). In T, the second record is a near miss. TC adds a completeness record stating that no other qualifying record exists. TX adds an ordinary record that is not a completeness record (in every family, a cover or status email about the same matter). TS adds an exclusion whose own words leave a qualifying route open. Under an open record, K2 is PASS, TC is FAIL, and T, TX, and TS are HOLD; under a closed record, only K2 is PASS.

3.2 Two completion label certificates

As in Round 44, a label is HOLD under an open record exactly when the packet admits two consistent completions (added records not contradicted by the packet) with different verdicts. Each family's certificate enumerates the completion events the packet leaves open; an exclusion removes a completion event only if that event lies inside its stated scope. TC's exclusion removes every completion; TS's leaves at least one.

3.3 The record set ledger and four rule sets

The ledger asks, for each requirement, for a status per packet record (SATISFIES, PART, FAILS, or IRRELEVANT), for record sets (groups of quotes, one from each record, that together meet every part), and for an exclusion field (quote, stated by, excludes, period, covers_requirement). The witness does not decide the audit question. Deterministic code checks every quote against the packet and resolves:

Rule set	Record sets	Scope test	Locality test	Status
R44	no	no	no	Round 44 rules: MET only on a SATISFIES record; any verified exclusion gives UNMET
R44+S	yes	no	no	ablation
R44+X	no	yes	no	ablation
R45	yes	yes	no	pre registered primary rule
R45+L	yes	yes	yes	exploratory, declared before any call

A record set counts only if every quote verifies and the quotes lie in at least two different records. Under the scope test, an exclusion counts only if its quote verifies, lies in a record, and covers_requirement is true; otherwise the condition is UNDETERMINED. Under the locality test, an exclusion quote that lies in a record the ledger marks FAILS or PART is read as that record's own statement and ignored. The open resolver maps any UNDETERMINED condition to HOLD; the closed resolver reads it as UNMET.

3.4 Exhibit track with annotated ledgers

Round 44's gated rewrite made no wrong rewrite because none was invited. Round 45 adds, for each exhibit, an annotated ledger: the correct gold ledger plus a reviewer note that is wrong. On HOLD exhibits the note says the instrument was filed in executed form, so its conditions were met; on CERTIFY exhibits the note says the ledger omits a condition and quotes a clause that is not one. A rewrite that follows the note is wrong in each case. The validator prompts are unchanged from Round 44 except for one sentence telling validators that ledger notes are unverified.

3.5 Run order, budget, and completion

The run plan was fixed before any call (`run-plan-r45.md`): blind review by Witness A at maximum reasoning (two calls of three families); record set ledgers by Witnesses B and C on all 30 items; open declaration direct readings by Witnesses C and B on all 30 items; Witness C on the 24 exhibits (rewrite of the seeded ledger, rewrite of the annotated ledger, flag only on the annotated ledger); record set ledgers by Witness D until a sub cap; and one Witness A synthesis call. If the budget stops a phase, every later call is skipped and reported.

The review phase used a larger share of the budget than planned, and the record set phase used more per call for Witness B than in Round 44. The budget rule first stopped the round after 57 of 60 record set calls, skipping 3 record set calls (Witness C on `M06TX` and `M06TS`, Witness B on `M06TS`), all 60 direct readings, all 72 exhibit validator calls, all 30 Witness D calls, and the synthesis call.

Same day completion (post hoc). Later on November 3, 2026, after the first results had been read and written up, the skipped calls were run with the same prompts, settings, witness assignments, and phase order, under a new run budget large enough for every remaining call. The loader reuses every completed call from the transcript, so no completed call was repeated; the 3 budget stopped records remain in the transcript. All phases completed: 3 record set calls, 60 direct readings, 72 exhibit validator calls, 30 Witness D calls (no sub cap bound), and the synthesis call. Because the decision to complete the round was made after the first results were seen, the completed phases should be read as same day, post hoc measurements rather than as part of a single uninterrupted pre registered run. The hypotheses, rule sets, scoring code for the first phases, and outcome measures were not changed. Scoring code for the direct reader and exhibit phases was written after the rerun, following the outcome measures in the run plan. One analysis fix was made after the synthesis call: the R44 against R45 flip table had pooled Witness D's ledgers with Witnesses B and C; it now covers Witnesses B and C only, as the plan specifies, and Witness D is reported separately (Section 7.6).

---

4. Related work

Fact verification datasets already treat evidence as sets: FEVER labels a claim with one or more evidence sets, each of which may need several sentences together [S61], and HotpotQA asks for supporting facts drawn from more than one document [S62]. The record set field brings that structure into a contract ledger and lets the resolver check it. Separating a reader that translates text into a formal structure from a solver that decides follows Faithful Chain of Thought [S63] and Logic-LM [S64]; Round 45 keeps the translation step and studies where the solver's fixed rules are wrong. The question whether a fixed rule or a case by case judgment should decide is the rules versus standards question of legal theory, where rules are cheap to apply and wrong at the edges and standards are costly and contextual [S65]. Argumentation frameworks treat an exclusion as an attack that defeats an argument only if it is itself acceptable [S66]; the scope test is a narrow version of that idea. The open and closed record conventions are the open and closed world assumptions [S49]. The misleading reviewer note tests whether a validator defers to an upstream claim, a behavior studied as sycophancy [S67]. Label certificates and blind review follow Round 44, motivated by evidence that label errors in test sets change conclusions [S57]; using a model reviewer follows the judge literature [S58]. Contract reading benchmarks include ContractNLI [S53]; contrast variants follow Gardner et al. [S40]; abstention is surveyed by Wen et al. [S24]; zero event limits use the exact binomial limit [S47]; exhibits come from EDGAR [S30].

---

5. Hypotheses

Fixed before any call.

ID	Statement	Test
H45.1	On the same ledgers, R45 under the open resolver gives more correct outcomes than R44 without more unsafe releases.	E45-B, rule comparison on rs ledgers
H45.2	Record sets recover K2: R45 resolves K2 to PASS where R44 cannot.	E45-B, K2 rows
H45.3	The scope test removes unsafe FAILs on TS under the open resolver.	E45-B, TS rows
H45.4	The completeness content, not an added record, drives FAIL: TC resolves FAIL far more often than TX.	E45-B, TC against TX
H45.5	The exploratory locality rule improves on R45 without new errors.	E45-B, R45+L against R45
H45.6	Where the resolver is wrong, the same witness reading directly is right more often than wrong.	E45-B, rule against reader two by two
H45.7	A rewriting validator follows a misleading reviewer note on some annotated ledgers, and the unchanged gate catches the resulting releases but not the resulting holds.	E45-C

---

6. Evidence category 1: deterministic diagnostic (`E45-A`)

No model calls. Script: `run_structural_diagnostic_r45.py`; output: `structural_results.json`.

Certificate	Result
S1 no single record meets the requirement in any K2 item	6 of 6 families
S1 TC, TX, and TS differ from T by exactly one added record	6 of 6 families
S1 completeness exclusion covers the open completion; scoped exclusion does not	6 of 6 families
S1 labels equal the certified labels (K2 PASS and PASS; T, TX, TS HOLD and FAIL; TC FAIL and FAIL)	30 of 30 items
S2 gold ledger quotes verify against the packets	30 of 30
S4 exhibits new against Rounds 39 to 44	24 of 24 accessions and filers

Gold ledgers under each rule set (correct items per variant out of 6; unsafe count over all 30):

Rule set	Resolver	K2	T	TC	TX	TS	Unsafe
R44	open	0	6	6	6	0	6
R44	closed	0	6	6	6	6	6
R44+S	open	6	6	6	6	0	6
R44+S	closed	6	6	6	6	6	0
R44+X	open	0	6	6	6	6	0
R44+X	closed	0	6	6	6	6	6
R45	open and closed	6	6	6	6	6	0

Under the Round 44 rules, a correct gold ledger for K2 can never reach PASS (it holds under the open resolver and fails unsafely under the closed one), and a correct gold ledger for TS fails unsafely under the open resolver because the scoped exclusion is read as covering everything. These are rule failures by construction: the ledger is right and the rule is wrong.

Scripted reader mistakes (each derived from a gold ledger by one change; open resolver; correct out of n):

Reader mistake	n	R44	R44+S	R44+X	R45	R45+L
record set omitted (K2)	6	0	0	0	0	0
record scoped negative quoted as exclusion, covers false (T, TX, TS)	18	0	0	18	18	18
record scoped negative quoted as exclusion, covers true (T, TX, TS)	18	0	0	0	0	18
scope gap missed, covers true (TS)	6	0	0	0	0	0
added record quoted as exclusion, covers false (TX)	6	0	0	6	6	6
completeness scope denied, covers false (TC)	6	6	6	0	0	0

Every rule set is defeated by some reader mistake, and each mistake defeats a different set of rules. No rule set can recover a record set the reader does not list or a scope the reader misjudges as covering. The scope test makes the resolver depend on the reader's covers_requirement judgment in both directions: it rescues an unscoped negative when the reader says it does not cover, and it loses a

completeness record when the reader says it does not cover. The locality test catches the one mistake that defeats every pre registered rule.

Exhibit ledgers. Gold ledgers resolve correctly on 24 of 24 exhibits; seeded ledgers resolve wrongly on 24 of 24; annotated ledgers left unchanged resolve correctly on 24 of 24 (the resolver ignores the note). A rewrite that follows the note releases on all 12 HOLD exhibits and holds on all 12 CERTIFY exhibits. The unchanged gate routes all 12 note following releases to review and accepts all 12 note following holds. A correct repair of a seeded ledger is routed to review on the 12 seeded CERTIFY ledgers, as in Round 44.

Zero event limits (exact one sided 95 percent upper limit): 0.393 at 6, 0.221 at 12, 0.153 at 18, 0.117 at 24, 0.095 at 30, 0.049 at 60.

---

7. Evidence category 2: live authored families (`E45-B`)

7.1 Blind label review

Witness A, at maximum reasoning, reviewed all 30 packets under opaque identifiers in two calls of 15 packets. Its open and closed verdicts agreed with the authored labels on all 30 packets (60 of 60 verdicts). It reported 7 defect notes:

Item	Defect noted by the reviewer (paraphrased)
`M01TS`	the agent's certificate excludes only Buyer counterparts and may be misread as a complete exclusion
`M01TX`	the agent's email says it will process the instruction, which a reader could argue is a waiver
`M02K2`	the director's written consent postdates the board approval
`M03TS`	the courier's statement could be misread as excluding all delivery routes
`M04TC`	the register lists bonds by number without sums, though a careful reader would still reach FAIL
`M05K2`	the report's freshness test runs from a scheduled funding date
`M06K2`	the 75 percent threshold is met with no margin on a rent roll three months old

Two of the notes (`M01TS`, `M03TS`) describe the intended difficulty of the TS variant rather than a defect. Under the fixed adjudication rule, a reported defect marks the family contested without changing any packet. Every family received at least one note, so every family is contested, and the uncontested only analysis is empty. All results below are on all six families.

7.2 Record set ledgers under the four rule sets

Witnesses B and C returned 60 of 60 ledgers. Cells are correct / unsafe / other wrong, by witness:

Rule set	Resolver	K2 B	K2 C	T B	T C	TC B	TC C	TX B	TX C	TS B	TS C
R44	open	1/0/5	1/0/5	4/2/0	6/0/0	3/0/3	6/0/0	4/2/0	6/0/0	3/3/0	2/4/0
R44+S	open	5/0/1	6/0/0	2/4/0	6/0/0	4/1/1	6/0/0	0/6/0	6/0/0	2/4/0	2/4/0
R44+X	open	1/0/5	1/0/5	6/0/0	6/0/0	2/0/4	6/0/0	6/0/0	6/0/0	6/0/0	6/0/0

Rule set	Resolver	K2 B	K2 C	T B	T C	TC B	TC C	TX B	TX C	TS B	TS C
R45	open	5/0/1	6/0/0	5/1/0	6/0/0	2/1/3	6/0/0	5/1/0	6/0/0	5/1/0	6/0/0
R45	closed	5/0/1	6/0/0	5/1/0	6/0/0	4/1/1	6/0/0	5/1/0	6/0/0	4/1/1	5/0/1
R45+L	open	5/0/1	6/0/0	5/1/0	6/0/0	0/1/5	6/0/0	5/1/0	6/0/0	5/1/0	6/0/0

Totals over the 60 ledgers:

Rule set, resolver	Correct	Unsafe	Other wrong	Unsafe upper 95 percent
R44, open	36	11	13	0.285
R44+S, open	39	19	2	0.429
R44+X, open	46	0	14	0.049
R45, open (primary)	52	4	4	0.146
R45, closed	52	4	4	0.146
R45+L, open (exploratory)	50	4	6	0.146
R45+F, open (post hoc)	55	0	5	0.049

By witness under R45 with the open resolver: Witness C 30 of 30 correct, 0 unsafe (upper limit 0.095); Witness B 22 of 30 correct, 4 unsafe, 4 holds where a decision was due. The 3 ledgers completed in the same day rerun (Witness C on `M06TX` and `M06TS`, Witness B on `M06TS`) added 2 correct ledgers for Witness C and 1 unsafe release for Witness B, of the same kind as Section 7.5.

7.3 Where the fixed rule was wrong and the witness was right

Rule against rule. Comparing R44 and R45 on the same ledger isolates the effect of the rules, because the witness output is identical:

	R45 right	R45 wrong
R44 right	33	3
R44 wrong	19	5

All 19 rescues are cases where the Round 44 rule was wrong on a ledger whose relevant content was right. In 9 (K2, both witnesses, every family except `M05`) the witness listed a valid record set that R44 cannot credit. In 10 (T, TX, and TS) the witness quoted a negative statement and itself marked it as not covering the requirement, and R44 read it as an exclusion anyway. Two K2 ledgers (Witness B on `M02K2`, Witness C on `M05K2`) were right under R44 for a different reason: the witness split the requirement into parts that one record each satisfies. Splitting a requirement is the other way to express joint satisfaction, and it is the same move that produced the `M06` failure below.

The 3 reverse cases are all Witness B. On `M02TC` the witness marked the completeness record's exclusion as not covering the requirement, so the scope test held where R44's broader rule failed correctly. On `M06T` and `M06TX` the new rule released an unsafe PASS (Section 7.5). The 5 ledgers wrong under both are Witness B: two with a condition quote that did not verify (`M01TC`, `M05K2`), one where the witness judged a completeness exclusion out of scope (`M04TC`), and two on the estoppel family (`M06TC`, `M06TS`) where a record set containing the stale certificate was credited.

Rule against reader. The same two witnesses also read every packet directly under the unchanged open declaration (DO). Direct readings were correct on 46 of 60 (Witness B 22, Witness C 24), with 14 unsafe FAILs and no false holds. By variant: K2 12 of 12, TC 12 of 12, T 3 of 12, TX 10 of 12, TS 9 of 12. Every direct reader error was a FAIL on an item whose open label is HOLD, the near miss error that Rounds 43 and 44 measured.

Per item and witness, the R45 open resolver on that witness's ledger against that witness's direct reading:

	Reader right	Reader wrong
Resolver right	39	13
Resolver wrong	7	1

The 7 resolver failures (resolver wrong, reader right) are all Witness B: 3 where a record set credited a record the same ledger marks FAILS (`M06TX`, `M06TC`, `M06TS`), 2 where the witness judged a completeness exclusion out of scope (`M02TC`, `M04TC`), and 2 where a condition quote did not verify (`M01TC`, `M05K2`). The one ledger wrong under both is Witness B on `M06T`, where the resolver released PASS and the reader returned FAIL. The 13 cases where the resolver was right and the reader wrong are near miss items (T, TX, TS) read directly as FAIL.

7.4 Completeness record against added record

Under R45 with the open resolver:

Variant	Ledgers	FAIL	HOLD	PASS
TC (completeness record)	12	8	3	1
TX (added record, not completeness)	12	0	11	1
TS (scoped exclusion)	12	0	11	1
T (near miss only)	12	0	11	1

No ledger on TX, TS, or T resolved to FAIL. FAIL appeared only where the added record was a completeness record. The 3 TC holds are Witness B ledgers in which the witness judged the completeness record not to cover the requirement or a quote failed to verify. The single PASS in each row is Witness B on `M06` (Section 7.5).

Direct readings of the same items: TC FAIL on 12 of 12, TX on 2 of 12, TS on 3 of 12, and T on 9 of 12. The direct reader's FAIL rate on TX (2 of 12) is lower than on T (9 of 12), so for direct readers an added record that is not a completeness record lowered FAIL rather than raising it; this was not predicted and is reported as measured.

7.5 Failure walk-through: `M06T`, Witness B

The packet: a purchase agreement requires that, before the June 30 closing, the seller deliver "estoppel certificates signed by tenants that together lease at least 75 percent of the rentable area of the Property, each certificate dated no earlier than 30 days before the Closing Date." REC-1 is a rent roll (Alder 45 percent, Birchway 30 percent). REC-2 is Alder's certificate dated June 12. REC-3 is Birchway's certificate "dated May 10, 2026 … Bring down certificate: none attached." The label is HOLD

under an open record (a later Birchway certificate, or a certificate from Cedar or Dunmore, may exist) and FAIL under a closed record.

Witness B split the requirement into four conditions: delivery, tenant signatures, rentable area threshold, and certificate date limit. Every record status it gave was correct. Under the date limit it marked Alder SATISFIES and Birchway FAILS ("May 10, 2026 is earlier than 30 days before the June 30, 2026 Closing Date"). Under the threshold it listed a record set of the rent roll, Alder's certificate, and Birchway's certificate ("45 percent plus 30 percent equals 75 percent"). It quoted "Bring down certificate: none attached." as an exclusion and marked it as not covering the requirement.

R45 then resolved each condition separately. The date limit was MET because one record satisfied it; the threshold was MET because the record set verified; the exclusion was out of scope. Every condition was MET, so R45 released PASS. Witness C, reading the same packet as one condition, marked Birchway FAILS and listed no set, and R45 held.

The witness was right about every record. The rule was wrong twice: it read "each certificate dated within 30 days" as satisfied by any one certificate, and it credited a record set built from a record the same ledger had just marked FAILS. R44 held this item only because it cannot credit record sets at all. The same pattern produced Witness B's unsafe PASS on `M06TX`, `M06TC`, and `M06TS`.

7.6 Witness D panel

Witness D wrote record set ledgers on all 30 items, one call at a time in opaque identifier order; the phase sub cap did not bind. Under R45 with the open resolver Witness D was correct on 22 of 30 with 0 unsafe releases (upper limit 0.095) and 8 holds where a decision was due: all 6 TC items (each with an undetermined condition, the witness did not treat the completeness record as resolving the requirement) and 2 K2 items (`M03K2`, `M06K2`). T, TX, and TS were correct on 18 of 18. Under the Round 44 rules the same ledgers gave 19 correct and 5 unsafe. Witness D is a separate panel and is not pooled with Witnesses B and C.

---

8. Evidence category 3: real SEC exhibits (`E45-C`)

8.1 Track

24 exhibits, fixed in advance, each with a gold ledger, a seeded ledger (span valid near miss evidence marked SATISFIED on HOLD exhibits; a spurious condition on CERTIFY exhibits), and an annotated ledger (the gold ledger with a misleading reviewer note). Strata: HC, an effectiveness section listing conditions precedent; HA, conditions plus a statement that the instrument is effective as of its date; HR, receipt of counterparts as the only condition; CF, effective on its face; CD, effective on its face with distractors (counterparts mechanics, termination rights, conditions of another agreement, ratification clauses).

ID	Filer	Exhibit	Filed	Accession	Stratum	Label
`X01`	Barings Capital Investment Corp	EX-10.2	2026-10-27	`0001811972-26-000020`	HC	HOLD
`X02`	NexPoint Real Estate Finance, Inc.	EX-10.2	2026-05-15	`0001193125-26-226577`	CF	CERTIFY

ID	Filer	Exhibit	Filed	Accession	Stratum	Label
X03	Phillip Street BDC LLC	EX-10.1	2026-08-11	0001193125-26-344614	HC	HOLD
X04	United States Natural Gas Fund, LP	EX-10.16	2026-10-15	0002071876-26-000249	CF	CERTIFY
X05	Viemed Healthcare, Inc.	EX-10.1	2025-11-07	0001729149-25-000151	HC	HOLD
X06	Chimera Investment Corp	EX-2.2	2025-08-06	0001628280-25-038345	CF	CERTIFY
X07	Energy Recovery, Inc.	EX-10.1	2026-05-06	0001421517-26-000041	HC	HOLD
X08	Affirm Holdings, Inc.	EX-10.1	2024-02-08	0001820953-24-000006	CF	CERTIFY
X09	Vintage Wine Estates, Inc.	EX-10.4	2024-05-15	0000950170-24-060487	HC	HOLD
X10	Vertiv Holdings Co	EX-2.4	2026-02-13	0001674101-26-000008	CF	CERTIFY
X11	Enova International, Inc.	EX-10.47	2026-02-20	0001193125-26-060461	HC	HOLD
X12	Greif, Inc.	EX-10.1	2025-08-28	0000043920-25-000048	CF	CERTIFY
X13	New Mountain Finance Corp	EX-10.1	2025-05-05	0001496099-25-000018	HA	HOLD
X14	RE/MAX Holdings, Inc.	EX-10.4	2026-04-28	0001104659-26-049628	CD	CERTIFY
X15	Fox Factory Holding Corp	EX-10.1	2024-11-01	0001424929-24-000128	HA	HOLD
X16	Amplify Energy Corp.	EX-2.1	2025-04-15	0001104659-25-035119	CD	CERTIFY
X17	Applied UV, Inc.	EX-10.3	2024-04-02	0001213900-24-029453	HA	HOLD
X18	Signing Day Sports, Inc.	EX-10.1	2024-11-06	0001213900-24-094996	CD	CERTIFY
X19	Cable One, Inc.	EX-10.1	2025-05-02	0001632127-25-000064	HR	HOLD
X20	CION Investment Corp	EX-10.1	2024-06-18	0001104659-24-072879	HA	HOLD
X21	National Beverage Corp	EX-10.17	2025-07-02	0001437749-25-022007	HR	HOLD
X22	WhiteHorse Finance, Inc.	EX-10.1	2026-03-13	0001104659-26-027685	CD	CERTIFY
X23	Direct Selling Acquisition Corp.	EX-2.1	2025-01-03	0001193125-25-001570	CD	CERTIFY
X24	AppLovin Corp	EX-2.1	2025-07-01	0001751008-25-000061	CD	CERTIFY

Each exhibit's filing index and document links are in `edgar_track_r45.json`. One stratum assignment changed during track construction, before any call: `X20` was first drawn as a face effective candidate, but its Section 3 makes effectiveness conditional on payment of fees, so it was relabeled HOLD (HA) and two face effective exhibits (`X23`, `X24`) were added to keep 12 per label.

8.2 Validator results

Witness C made 72 calls, all completed. The rewritten ledger was resolved by the unchanged exhibit resolver (fail closed) and passed through the unchanged gate.

Arm	HOLD exhibits (12)	CERTIFY exhibits (12)	Gate
Rewrite of seeded ledger	12 correct	12 correct	12 HOLD accepted; 12 correct CERTIFY repairs routed to review

Arm	HOLD exhibits (12)	CERTIFY exhibits (12)	Gate
Rewrite of annotated ledger	12 correct	12 correct	all 24 accepted (no change of outcome)
Flag only, annotated ledger	12 correct (10 flagged)	12 correct (0 flagged)	not applicable

No rewrite followed the misleading reviewer note: on 0 of 24 annotated ledgers did the rewritten ledger change the resolver outcome, so there was no note following release for the gate to catch and no note following hold for it to miss. The flag only check flagged 10 of 12 annotated HOLD ledgers, on which the note urged a release, and none of the 12 CERTIFY ledgers, on which the note urged a hold; flags on HOLD ledgers leave the decision at HOLD, so they cost nothing here. The gate's only cost on this track was routing the 12 correct repairs of the seeded CERTIFY ledgers to review, as in Round 44. Zero event limits: 0 of 24 note following rewrites, upper 95 percent limit 0.117; 0 of 72 wrong outcomes, upper limit 0.041.

---

9. Findings and interpretations

9.1 Measured findings

- F45.1 (deterministic, E45-A): on correct gold ledgers, the Round 44 rules are wrong on all 6 K2 items under both resolvers and release an unsafe FAIL on all 6 TS items under the open resolver; R45 reproduces all 60 labels.
- F45.2 (deterministic, E45-A): each scripted reader mistake defeats a different set of rules; a record scoped negative quoted as an exclusion and marked as covering defeats all four pre registered rule sets and is caught only by the locality test.
- F45.3 (deterministic, E45-A): on 24 SEC exhibits, a rewrite that follows the misleading note releases on all 12 HOLD exhibits and holds on all 12 CERTIFY exhibits; the unchanged gate routes all 12 releases to review and accepts all 12 holds.
- F45.4 (live, E45-B): the blind review agreed with 60 of 60 labels and raised 7 defect notes, one or more in every family, so every family is contested under the fixed rule.
- F45.5 (live, E45-B): on 60 ledgers, R45 with the open resolver was correct on 52 with 4 unsafe releases, against 36 correct and 11 unsafe under the Round 44 rules on the same ledgers; Witness C was correct on 30 of 30.
- F45.6 (live, E45-B): on the same ledgers, R44 was wrong and R45 right on 19 (9 credited record sets, 10 unscoped negatives no longer read as FAIL), and R44 right and R45 wrong on 3, all Witness B.
- F45.7 (live, E45-B): under R45, TC resolved to FAIL on 8 of 12 ledgers and TX, TS, and T on 0 of 12 each.

- F45.8 (live, E45-B): record sets without the scope test raised unsafe releases from 11 to 19; the scope test without record sets removed all unsafe releases but held 10 of 12 K2 ledgers.
- F45.9 (live, E45-B): the exploratory locality rule changed no unsafe outcome and lost 2 TC ledgers where Witness B marked the completeness record itself FAILS.
- F45.10 (live, E45-B): direct readers were correct on 46 of 60 with 14 unsafe FAILs, all on near miss items; the resolver was wrong and the reader right on 7 ledgers, the resolver right and the reader wrong on 13, and both wrong on 1.
- F45.11 (live, E45-B): Witness D, as a separate panel, was correct on 22 of 30 under R45 with 0 unsafe releases and 8 false holds, including all 6 TC items.
- F45.12 (live, E45-C): Witness C's 72 exhibit validator calls produced 72 correct outcomes; no rewrite followed the misleading note (0 of 24) and the gate routed 12 correct repairs to review.
- F45.13 (completion): all phases after the first 57 record set calls were completed in a same day, post hoc rerun with unchanged prompts and settings.

9.2 Interpretation

The roadmap asked where fixed rules are wrong and witnesses are right. On these families the answer has three parts.

First, the Round 44 rules were wrong where Round 44 said they would be, and the two fields Round 44 proposed fix them when the witness fills them in. Most of the improvement from 36 to 52 correct is the rule catching up with what the witness already wrote. Witnesses also have a second way to express joint satisfaction, splitting a requirement into parts, and the resolver handles that way well only when the parts are truly independent.

Second, a rule change is not one fix but a new interaction. Record sets alone made things worse, because the old rule's holds on near miss items had partly come from an unrelated undetermined condition that masked an unscoped exclusion. Removing the mask exposed the exclusion. The two features are safe together and not apart, which argues for changing resolver rules only as tested combinations.

Third, the most interesting failure was not a reader failure. In `M06T` the witness gave every record the right status, and the rule combined those statuses wrongly, because it treats the conditions a witness lists as independent and existential. A requirement with "each" in it is universal over the records that count toward it. When a witness splits such a requirement, the resolver needs to keep the link between the parts. The post hoc consistency test (a record set may not use a record the same ledger marks FAILS) removed all 4 unsafe releases on these ledgers (55 of 60 correct, 0 unsafe), but it was written after reading the ledgers and must be tested fresh.

The completeness control is the clearest result. An added record of the same kind and length never produced FAIL; only a record that states completeness did. That rules out the reading that ledgers resolve TC items to FAIL simply because the packet got longer or because any record after the near miss prompts a conclusion.

The direction of reader dependence also matters for P2. The scope test makes the resolver trust the witness's covers_requirement judgment. That judgment rescued 10 ledgers and lost 3 completeness

ledgers. A resolver that defers to a witness judgment inherits that witness's errors; P2's guarantee is only as strong as the fields the code can check itself.

The rule against reader comparison answers the roadmap question in the other direction as well. Where the resolver was wrong, the same witness reading directly was usually right (7 of 8), but the reverse was nearly twice as common (13 cases where the resolver was right and the reader wrong), and the two kinds of error almost never coincided. That suggests a composition in which a disagreement between the resolver and a direct reading is a cheap trigger for review, rather than a choice of one over the other. This is an interpretation from one round on authored families.

The exhibit track did not produce the failure it was built to measure: the strongest validator ignored a misleading note every time. The track therefore says nothing yet about whether the gate catches note following; it does show that telling validators notes are unverified did not cost any correct outcome.

9.3 Hypothesis status

ID	Status	Basis
H45.1	Supported (52 of 60 correct and 4 unsafe under R45, against 36 and 11 under R44, on the same ledgers)	F45.5
H45.2	Supported (K2 correct 11 of 12 under R45, 2 of 12 under R44)	Section 7.2
H45.3	Supported (TS 11 of 12 held under R45; 7 of 12 unsafe under R44; the remaining TS release is the Section 7.5 failure)	Section 7.2
H45.4	Supported (TC FAIL 8 of 12, TX FAIL 0 of 12)	F45.7
H45.5	Not supported (no gain; 2 TC ledgers lost)	F45.9
H45.6	Supported, same day post hoc completion (where the resolver was wrong, the reader was right on 7 of 8)	F45.10, Section 3.5
H45.7	Not supported live (no rewrite followed the note, 0 of 24, so the gate had nothing to catch); structural part supported	F45.3, F45.12

10. Toward Theory v1

P2 (a deterministic resolver over extracted evidence) survives this round with named exceptions. The resolver is wrong, by construction, wherever its rules cannot express what a correct ledger says: joint satisfaction before record sets, scoped negation before the scope test, and universal requirements split into independent conditions even after both. Each exception is a rule failure, not a reader failure, and each was visible in the structural diagnostic or in a single failure walk-through before it was visible in the counts. We therefore restate P2: a resolver is safe to the extent that its rules are complete for the requirement forms in the domain and enforce consistency across the parts of a ledger; every field it takes on trust from a witness (covers_requirement here) brings that witness's error rate into the resolver.

For the composition question that leads to Round 50, this round suggests that the line between witness and resolver should follow what code can check. Record sets are checkable (quotes verify, records

differ). Scope is not, so it remains a witness judgment, and a second witness asked only about scope may be the cheapest way to make it checkable.

---

11. Method changes for later rounds

1. Pre register a consistency test for resolvers: a record set may not include a record the same ledger marks FAILS under any condition, and a requirement stated with "each" or "every" is not MET while any counted record FAILS.
2. Ask a separate witness for covers_requirement and resolve only when two witnesses agree on scope.
3. Run the direct reader phase before the cheaper ledger phases when the rule against reader comparison is the question, so that a budget stop cannot remove it; this round needed a same day completion run for that reason.
4. Size the label review from this round's measured share of the run budget rather than from earlier rounds.
5. Use annotated exhibit ledgers with weaker validators, or stronger notes, so that note following occurs and the gate's catch rate can be measured.

---

12. Synthesis review

Run in the same day completion. Witness A, at maximum reasoning, received only the recorded results and was asked to separate what the data show from what they suggest. Its output is recorded verbatim in `synthesis-review.json` and is interpretation, not evidence. In summary:

- It read the R45 gain on K2 and TS as consistent with the new fields repairing both Round 44 weak points for Witnesses B and C, while noting that TC did not improve and that Witness D held every TC item.
- It read resolver and reader errors as complementary: reader errors are FAIL calls on HOLD items, resolver errors are holds and stray releases from quotes, scope, and record sets.
- It noted that the direct reader's 12 of 12 FAIL on TC is confounded by its 9 of 12 FAIL on plain T, so the completeness mechanism is cleanest in the resolver results.
- It noted that the exhibit track was at ceiling and so cannot show whether the gate catches note following.
- It flagged an inconsistency in the brief it received: the R44 against R45 flip table pooled Witness D's ledgers with Witnesses B and C (90 comparisons). That was an analysis error, corrected after the call; the flip table in Section 7.3 covers Witnesses B and C only, as the run plan specifies. It also asked for the flag only denominator, now stated (10 of 12 HOLD, 0 of 12 CERTIFY).

- Its proposed next check: pre register R45+F against R45 on new families by other authors, resolve every panel under both rule sets, and build annotated exhibit ledgers on which a weaker validator demonstrably follows the note.

---

13. Limitations and threats to validity

- Six authored families, 30 items, and 60 ledgers from two witnesses. Witness C was correct on every ledger, so most of the variation comes from one witness; Witness D is a third panel with a different error profile.
- The run budget first stopped the round after 57 of 60 record set calls. The remaining phases were completed the same day, after the first results had been read, so they are post hoc in timing even though prompts, settings, and order were unchanged. Scoring code for the direct reader and exhibit phases was written after the rerun.
- Every family was marked contested by the review rule. Two of the seven notes describe the intended TS difficulty rather than defects; we report all families together and do not reclassify.
- The four rule sets were designed by the authors from Round 44's failures. R45 reproduces all gold labels by construction; its live advantage is a measure of how often witnesses fill the new fields correctly, not of rule quality in general.
- The consistency test (R45+F) was written after reading the ledgers and is reported only as a hypothesis.
- The same witnesses wrote ledgers and direct readings, so the rule against reader comparison is not between independent judges.
- Unsafe counts are small and the upper limits wide (0.146 at 4 of 60).
- The exhibit track was at ceiling for Witness C (72 of 72), so it does not measure the gate's catch rate on note following.

---

14. Next check

1. Rerun the record set ledger with a pre registered consistency test on new families that state "each" and "every" requirements, and measure whether unsafe releases from split conditions fall to zero without new holds.
2. Pre register the resolver against reader disagreement as a review trigger and measure its review cost and the errors it catches on new families.
3. Rerun the annotated exhibit ledgers with a weaker validator, or with notes that cite plausible window text, so that note following occurs and the gate's catch rate can be measured.

---

15. References

- [S24] Bingbing Wen, Jihan Yao, Shangbin Feng, et al. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, 2024.
- [S30] U.S. Securities and Exchange Commission. EDGAR Full-Text Search and Archives. https://efts.sec.gov/LATEST/search-index and https://www.sec.gov/Archives/edgar/data/.
- [S40] Matt Gardner, Yoav Artzi, Victoria Basmova, et al. Evaluating Models' Local Decision Boundaries via Contrast Sets. arXiv:2004.02709, 2020.
- [S47] James A. Hanley, Abby Lippman-Hand. If Nothing Goes Wrong, Is Everything All Right? Interpreting Zero Numerators. JAMA 249(13), 1983. doi:10.1001/jama.1983.03330370053031.
- [S49] Raymond Reiter. On Closed World Data Bases. In Hervé Gallaire and Jack Minker (eds.), Logic and Data Bases, Plenum Press, 1978, pp. 55 to 76. doi:10.1007/978-1-4684-3384-5_3.
- [S53] Yuta Koreeda, Christopher D. Manning. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. Findings of EMNLP 2021. arXiv:2110.01799.
- [S57] Curtis G. Northcutt, Anish Athalye, Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. NeurIPS 2021 Datasets and Benchmarks Track. arXiv:2103.14749.
- [S58] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track. arXiv:2306.05685.
- [S61] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. NAACL 2018. arXiv:1803.05355.
- [S62] Zhilin Yang, Peng Qi, Saizheng Zhang, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. EMNLP 2018. arXiv:1809.09600.
- [S63] Qing Lyu, Shreya Havaldar, Adam Stein, et al. Faithful Chain-of-Thought Reasoning. IJCNLP-AACL 2023. arXiv:2301.13379.
- [S64] Liangming Pan, Alon Albalak, Xinyi Wang, William Yang Wang. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. Findings of EMNLP 2023. arXiv:2305.12295.
- [S65] Louis Kaplow. Rules versus Standards: An Economic Analysis. Duke Law Journal 42(3), 1992, pp. 557 to 629. doi:10.2307/1372840.
- [S66] Phan Minh Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence 77(2), 1995, pp. 321 to 357. doi:10.1016/0004-3702(94)00041-X.
- [S67] Mrinank Sharma, Meg Tong, Tomasz Korbak, et al. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548, 2023.