

Encevra Research Round 40: Composition Before Capability: Six Noetis Composition Topologies Compared on Release Safety, Abstention, and Calls per Decision

Date: 2026-10-29

Research lab: Encevra Research

Short code: NCT-40 (Noetis Composition Topologies)

Primary tracks: Collaborative track ("When does dividing reasoning across specialists help, and when does it fragment evidence?"), Track 5 (Calibration, Abstention, and Release Control)

Position in the program: last round of the published "Ablation and memory" block (Rounds 26 to 40) and the first round written explicitly as input to the Round 41 to 50 "Theory v1" block.

Carry-forward boundary: Round 39 (CEV-39) held the composition of its witness panel fixed (a 3-of-4 quorum) and varied evidentiary lenses. It found that no witness abstained under lensed prompts. Round 40 holds the evidence questions fixed, removes lenses, and varies how witnesses are composed into one decision.

Abstract

Noetis is described on the Encevra site as a system of roles, evidence location rules, claim review, panel composition, synthesis, and a release gate. Rounds 35 to 39 each added one verification layer to that stack. None asked which way of composing the witnesses should carry the stack. Round 40 asks that question directly. We compare six composition topologies that release, hold, or route a contract audit decision: a single witness (T1), a forced majority of four witnesses (T2), one evidence extraction followed by a deterministic, span checked resolver (T3), a role specialized pipeline of extractor, validator, and decider (T4), a critique and revision loop (T5), and a disagreement routed hybrid that runs two independent extractions through the same resolver and sends any disagreement to review (T6). The metrics are accuracy, unsafe releases, abstention calibration (correct holds versus false holds), and model calls per decision.

We report three strictly separated evidence categories.

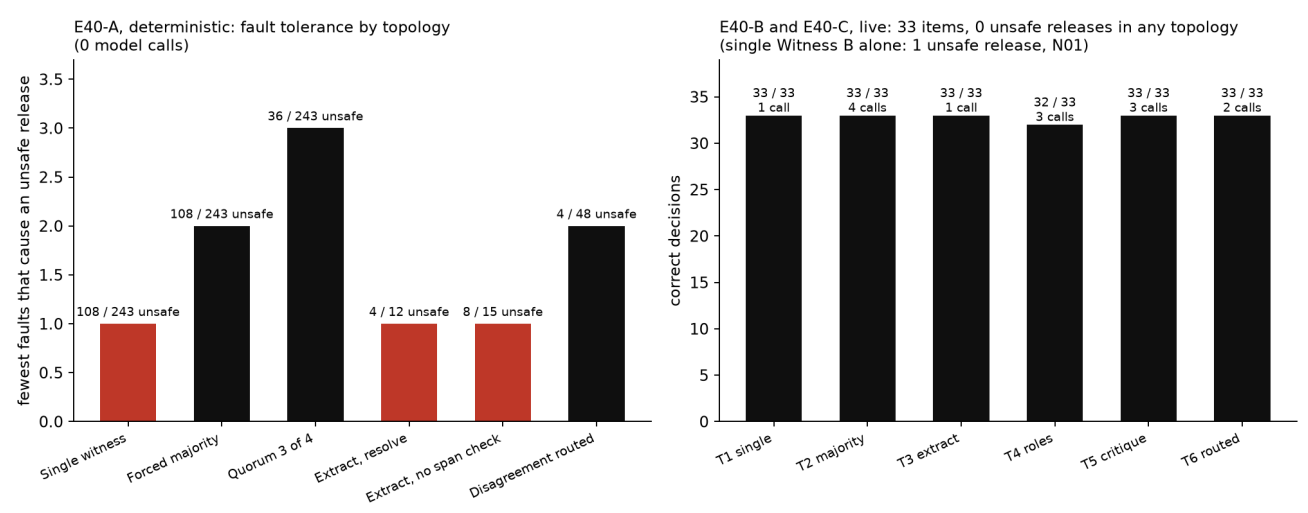
Deterministic (E40-A, no model calls). An exact enumeration of 243 verdict states shows that a single witness and a forced majority each release 108 / 243 states unsafely; a forced majority raises the fewest faults needed for an unsafe release from 1 to 2 but does not lower the unsafe count. When witnesses never return HOLD (the Round 39 behavior), neither topology holds any of 16 states whose truth is HOLD, and even a 3-of-4 quorum releases 10 / 16. For extraction topologies, a single extraction with a span check releases 4 / 12 fault states unsafely and needs only 1 fault; removing the span check raises this to 8 / 15. Routing extractor disagreement to review raises the minimum to 2 faults and limits unsafe releases to the 4 / 48 states in which both extractors make the identical misreading.

Live authored (E40-B, 17 packets). Seventeen frozen packets (eleven Round 39 packets and six new ones, 5 with truth HOLD) were run through all six topologies, built from ten shared calls per packet. Every topology scored 17 / 17 with 0 unsafe releases. Without a lens instruction, all four witnesses

returned HOLD on K08, the packet on which all eight lensed runs in Round 39 returned a determinate verdict. One single witness failure occurred: Witness B released PASS on N01, two same-date lender letters that grant and decline one approval. The majority, the extraction topologies, and the critique loop all removed it.

Live real exhibits (E40-C, 16 SEC EDGAR exhibits). On sixteen real exhibits from the Round 39 track (12 conditional, 4 effective on their face), five topologies scored 16 / 16. The role specialized pipeline scored 15 / 16: on X24 the validator rewrote the extractor's correct ledger by adding five ordinary covenants as "conditions", and the decider then held an instrument that is effective on its face.

Interpretation, kept separate. On this pool the witnesses were strong enough that composition did not change accuracy, so the live data cannot rank topologies by accuracy. What separated them was fault tolerance and the number of sequential model stages that can introduce a new error. The one live unsafe single answer was caught by every topology with more than one independent look, and the one live composed error was introduced by an added model stage. This supports a structure for Noetis in which models read and quote, a deterministic resolver decides, and a second independent read is spent on routing disagreement, not on voting.



1. Archive continuity and anti-repetition verification

Before designing any Round 40 experiment we rebuilt the archive ledger from the live research registry and the Round 39 manuscript (prior-rounds-ledger.json, 33 entries). A keyword scan of registry titles and questions for the Round 40 constructs (composition topology, majority vote, extract then resolve, deterministic resolve, role specialized, disagreement routed or routing, critique and revise, calls per decision, correlated misreading) returned 0 hits in Rounds 1 to 39. The registry indexes titles and questions only, so the scan cannot rule out a passing mention inside an earlier manuscript body; we therefore also compared against the three closest rounds by hand.

Prior round	Focus	How Round 40 differs
Round 10, E10-A (proposed, never run)	Six interaction arms, including debate, testing whether exchanging evidence rather than conclusions makes corrections reusable (H10)	Round 40 runs a topology comparison rather than proposing one, and its endpoint is release safety and abstention per decision, not reuse of corrections
GTP-34 (Round 34)	Dependency topologies among verification steps	"Topology" here means how witnesses are composed into a decision, not the dependency graph of checks
CEV-39 (Round 39)	Conflicting collateral records under evidentiary lenses, quorum fixed	The quorum is no longer fixed; lenses are removed; six compositions are compared on the same items

Round 40 reuses eleven Round 39 packets and sixteen Round 39 exhibits as test items. The question asked of them is new. Removing the lens partially executes item 1 of the Round 39 handoff (a neutral abstention test); it does not satisfy that item's request for review of expected verdicts by someone outside the authoring process, and we do not claim it does.

2. Where this round sits in the published program

The research site's theory page is labelled "working draft v0". It states four working propositions: reliability is a property of the process rather than of any single model; locating a quotation and judging whether it supports a claim are separate operations; confidence belongs to the final revised answer; and an improvement counts only against a matched cost baseline on an adjudicated endpoint. It also states that the program's hypotheses are open and not yet tested as a theory.

The published roadmap divides the program into Rounds 1 to 15 (Foundations), 16 to 25 (Measured revision), 26 to 40 (Ablation and memory: "Isolate which components carry the effect; study working memory and procedure reuse"), and 41 to 50 (Theory v1: "Intended convergence into a first formal Theory of Collaborative Intelligence, keeping only hypotheses that survive testing"). The site marks the last block as a stated direction, not a commitment. Round 40 is an ablation of one component, panel composition, and is written so that its propositions can be carried into or dropped from Theory v1 by later tests. Nothing in this round is a theory; Section 10 lists propositions, each with the evidence for it and the test that could remove it.

3. Topologies and metrics

3.1 The six topologies

Code	Topology	What decides	Model calls	Sequential model stages
T1	Single witness	Witness A verdict	1	1
T2	Forced majority	Plurality of four direct verdicts, ties broken by Witness A	4	1

Code	Topology	What decides	Model calls	Sequential model stages
T3	Extract then resolve	Witness C extracts conditions, records, and verbatim quotes; a deterministic resolver checks every quote against the source and decides, falling closed	1	1
T4	Role specialized pipeline	Witness C extracts, Witness B validates and corrects the ledger, Witness C decides from the validated ledger	3	3
T5	Critique and revision	Witness B answers, Witness C critiques, Witness B revises	3	3
T6	Disagreement routed hybrid	Witness C and Witness B extract independently; the resolver runs on each; release only if both resolve to the same determinate verdict, otherwise route to review	2	1

Two reference compositions are also scored: the Round 38 and 39 3-of-4 quorum, and the deterministic resolver applied to the validated T4 ledger.

3.2 Metrics

A decision is a **correct release** (released verdict equals truth), an **unsafe release** (a determinate verdict that differs from truth, including any release when truth is HOLD), a **correct hold** (HOLD or review when truth is HOLD), or a **false hold** (HOLD or review when a determinate verdict is supportable). Accuracy is correct releases plus correct holds. Abstention calibration is read as correct holds against false holds. Cost is counted only as model calls per decision and sequential model stages. No other cost measure is used.

4. Related work

Majority voting over sampled answers improves reasoning accuracy [S09], and panels of diverse judges can outperform a single judge [S26]. Multiagent debate has been reported to improve factuality [S17], and debate was proposed as a safety mechanism [S31]. Iterative self feedback can improve outputs [S32], but intrinsic self correction without external feedback can also degrade reasoning [S33]. Compound systems that make more calls do not improve monotonically [S34]. The Condorcet jury result depends on independent voters, and correlation among voters weakens it [S35]; latent class models estimate observer error rates under conditional independence [S28]. Abstention in language models is surveyed in [S24]. Round 40 differs from these by scoring compositions on unsafe release and holds, with a deterministic, span checked resolver as one of the arms, on contract audit items whose truth includes HOLD.

5. Hypotheses (pre-stated before the live run)

- H40.1 A forced majority does not reduce unsafe releases when witnesses do not abstain, because it has no HOLD output that disagreement can produce.

- H40.2 Separating extraction from decision, with a span check, removes fabricated quotes as a route to release.
- H40.3 Routing extractor disagreement to review raises the minimum faults for an unsafe release from one to two, at the price of false holds when extractors disagree.
- H40.4 Each added sequential model stage is a new place for an error to enter, so pipelines and critique loops can introduce errors that the single stage did not make.
- H40.5 Without lens instructions, witnesses return HOLD on packets whose contract leaves a decisive conflict unresolved.

6. Evidence category 1: deterministic structural diagnostic (E40-A)

6.1 Verdict aggregation, 243 states

Truth is PASS, FAIL, or HOLD; each of four witnesses returns PASS, FAIL, or HOLD.

Topology	Unsafe	False hold	Correct	Fewest faults for an unsafe release
T1 single witness	108	54	81	1
T2 forced majority	108	54	81	2
3-of-4 quorum (reference)	36	126	81	3

Over a uniform state space the forced majority has the same unsafe count as one witness. It changes which states are unsafe: none needs fewer than two faulty witnesses. The quorum cuts unsafe releases to 36 by converting 72 more states into false holds.

6.2 Witnesses that never abstain, truth HOLD, 16 states

Topology	Holds	Releases
T1 single witness	0	16
T2 forced majority	0	16
3-of-4 quorum	6	10

This generalizes the Round 39 K08 result: if no witness can say HOLD, the only path to a hold is a deadlock, and a quorum deadlocks only on balanced splits.

6.3 Extraction topologies

Each extraction is correct, one of two misreadings that resolve to a specific wrong outcome, or an unverifiable field (a fabricated or omitted quote).

Topology	Fault states	Unsafe	False hold	Correct	Fewest faults for an unsafe release
T3 extract then resolve, span checked	12	4	4	4	1
T3 without span check	15	8	4	3	1

Topology	Fault states	Unsafe	False hold	Correct	Fewest faults for an unsafe release
T6 disagreement routed	48	4	28	16	2

All 4 unsafe T6 states are states in which both extractors make the same misreading. T6 is therefore exactly as safe as its extractors are decorrelated.

7. Evidence category 2: live authored packets (E40-B)

7.1 Protocol

All calls ran at temperature 0. Witness A ran at maximum reasoning; Witness B, Witness C, and Witness D ran at standard reasoning. Each item received ten calls: four direct answers, two independent extractions, one validation, one decision from the validated ledger, one critique, and one revision. All six topologies were assembled from these shared calls, so topologies are compared on identical model outputs where they share a stage. All 330 calls across E40-B and E40-C returned and parsed. Prompts offered HOLD with the instruction to use it only if no determinate verdict is supportable, and carried no lens.

7.2 Packets

Eleven Round 39 packets were reused (K07 was excluded because of the authoring defect disclosed in Round 39 Section 7.5). Six new packets were authored for this round:

Packet	Construction	Truth
N01	Rent reduction requires lender approval; two same-date letters signed by an "Authorized Signatory" of the lender grant and decline it	HOLD
N02	Licence extends to Germany if fiscal 2025 sales reach 2,000,000 units; the licensee's report says 2.1 million, the licensor's auditor says 1.94 million	HOLD
N03	A rate rises on board approval; two minutes extracts, both certified by the Corporate Secretary, conflict	HOLD
N04	Margin falls to 2.25 if leverage is below 3.00; two same-date CFO certificates state 2.85 and 3.10	HOLD
N05	As N01, but the lease makes a Credit Committee letter the sole evidence of approval, and that letter declines	FAIL
N06	As N02, but the sale under audit is in France, so the outcome does not depend on the conflict	PASS

7.3 Results

Topology	Correct release	Unsafe	Correct hold	False hold	Calls per decision
T1 single witness	12	0	5	0	1
T2 forced majority	12	0	5	0	4
T3 extract then resolve	12	0	5	0	1
T4 role pipeline	12	0	5	0	3
T5 critique and revision	12	0	5	0	3

Topology	Correct release	Unsafe	Correct hold	False hold	Calls per decision
T6 disagreement routed	12	0	5	0	2
3-of-4 quorum (reference)	12	0	5	0	4

Direct witness accuracy was 17 / 17 for Witness A, Witness C, and Witness D, and 16 / 17 for Witness B. The two extractions resolved to the same outcome on 17 / 17 packets, so T6 routed nothing to review.

7.4 Failure walk-through: N01

Witness B, answering alone, returned PASS and quoted the approval condition: "Paragraph 1 takes effect only upon the written approval of Saltmere Bank NA, given on or before June 15, 2026." Its rationale treated the granting letter as decisive and stated that "the packet provides no rule under which REC 2 nullifies that approval". That is the unsafe release the lens experiments of Round 39 produced, now without any lens. In T5, Witness C marked the answer unsound because the declining letter "directly contradicts REC 1 on the exact same date", and Witness B revised to HOLD. The other three witnesses held directly, so T2 held. Both extractions recorded the two letters, and the resolver held because the contract designates neither. The critique loop changed 1 of 17 answers, and that change was a correction.

7.5 K08 without a lens

In Round 39, all eight lensed runs on K08 returned a determinate verdict, and the verdict followed the lens. In Round 40, with the lens removed, all four witnesses returned HOLD. The prompts differ in more than the lens (Round 39 also asked for a governing record and a conflict flag), so this is an observation across rounds, not a controlled ablation.

8. Evidence category 3: real SEC EDGAR exhibits (E40-C)

8.1 Track

Sixteen exhibits from the frozen Round 39 track were used: X01 to X06 (amendments whose effectiveness is conditioned), X15 and X16 (partitioned effectiveness), X17 to X20 (merger agreements with conditioned consummation), and X21 to X24 (instruments effective on their face). The expected decision is CERTIFY for X21 to X24 and HOLD otherwise. Each item carries its accession number; dollar amounts in quoted filings are written as "USD" in the archive copies.

ID	Filer	Exhibit	Accession	Expected
X01	Waste Connections, Inc.	EX-10.1	0001104659-25-052526	HOLD
X02	Ameresco, Inc.	EX-10.1	0001628280-26-022433	HOLD
X03	Sonendo, Inc.	EX-10.1	0000950170-25-031237	HOLD
X04	Hamilton Insurance Group, Ltd.	EX-10.2	0001593275-25-000251	HOLD
X05	Spire Global, Inc.	EX-10.1	0000950170-25-038328	HOLD

ID	Filer	Exhibit	Accession	Expected
X06	Arq, Inc.	EX-10.1	0001515156-26-000008	HOLD
X15	Manhattan Bridge Capital, Inc.	EX-10.5	0001493152-25-027893	HOLD
X16	Local Bounti Corporation	EX-10.2	0001840780-25-000031	HOLD
X17	e.l.f. Beauty, Inc.	EX-2.1	0001600033-25-000037	HOLD
X18	TaskUs, Inc.	EX-2.1	0000950157-25-000390	HOLD
X19	Criteo S.A.	EX-2.1	0001628280-26-053414	HOLD
X20	First Seacoast Bancorp, Inc.	EX-2.1	0001193125-26-206551	HOLD
X21	BioXcel Therapeutics, Inc.	EX-10.1	0001104659-26-093710	CERTIFY
X22	Apollo Debt Solutions BDC	EX-10.1	0001193125-25-080186	CERTIFY
X23	Medalist Diversified REIT, Inc.	EX-10.3	0001104659-26-016379	CERTIFY
X24	Mammoth Energy Services, Inc.	EX-10.1	0001628280-25-034133	CERTIFY

8.2 Results

Topology	Correct release	Unsafe	Correct hold	False hold	Calls per decision
T1, T2, T3, T5, T6	4	0	12	0	1, 4, 1, 3, 2
T4 role pipeline	3	0	12	1 (X24)	3

All four witnesses answered all 16 exhibits correctly when answering alone, and the two extractions agreed on all 16.

8.3 Failure walk-through: X24

Both extractors returned no conditions and the quote "effective as of the date hereof". The validator, asked to correct any wrong or missing field, replaced the empty list with five account and reporting covenants from the consent and recorded the correction "Added the five stated conditions." The decider then held, citing a truncated audit window and "ongoing conditions". The validator also returned those conditions as plain strings rather than in the extraction schema, which caused the deterministic resolver to ignore them. The reference "resolver on the validated ledger" therefore scored X24 correctly only because of a schema mismatch; under a strict schema rule it would have failed closed to HOLD. We count that reference as unreliable on X24.

9. Findings and interpretations

9.1 Measured findings

1. **Deterministic.** A forced majority does not reduce unsafe releases over a uniform state space (108 versus 108 of 243); it raises the fewest faults needed from 1 to 2.

2. **Deterministic.** With non-abstaining witnesses, single and majority topologies hold 0 / 16 HOLD-truth states.
3. **Deterministic.** The span check halves unsafe single extraction outcomes (8 / 15 to 4 / 12). Disagreement routing raises the fault minimum to 2, and its only unsafe states are identical misreadings.
4. **Live.** On 33 items and 330 calls, every topology made 0 unsafe releases. Accuracy was 33 / 33 for T1, T2, T3, T5, and T6, and 32 / 33 for T4.
5. **Live.** One single witness unsafe answer occurred (Witness B on N01). Every multi-look topology removed it.
6. **Live.** The one composed error (T4 on X24) was introduced by an added model stage that changed a correct ledger.
7. **Live.** Without lens instructions, all four witnesses held on all 5 HOLD-truth authored packets except Witness B on N01, and on all 12 conditional exhibits.

9.2 Interpretation

The live pool did not discriminate topologies by accuracy: 32 of 33 items had identical direct verdicts from all four witnesses. The useful signal is in the two disagreements. N01 shows that a single independent look is the thing that fails, and that any second independent look (a vote, an extraction, a critique) can catch it. X24 shows that a second look which is allowed to rewrite evidence, rather than only to flag it, can introduce an error. Combined with E40-A, this suggests that the second call should be spent on an independent read whose only power is to cause a hold, which is what T6 does at 2 calls and 1 sequential stage. The forced majority spends 4 calls for the same live result and has no structural abstention. The critique loop worked here (1 correction, 0 damage) but has the same sequential, rewriting structure as T4, and the literature reports intrinsic self correction can harm [S33].

9.3 Hypothesis status

Hypothesis	Evidence	Status
H40.1	E40-A 6.1 and 6.2: forced majority 108 / 243 unsafe, same as one witness; 0 / 16 holds without abstaining witnesses	Supported
H40.2	E40-A 6.3: unverifiable quotes fail closed; no live fabricated quote occurred, so live data neither add nor remove support	Supported
H40.3	E40-A 6.3: fault minimum 2; live extractors never disagreed, so the false hold cost was not observed	Supported
H40.4	E40-C X24: the validation stage introduced the only composed error	Supported
H40.5	Witnesses held on 19 / 20 direct HOLD-truth authored evaluations and 48 / 48 conditional exhibit evaluations; Witness B released N01	Partly supported

10. Toward Theory v1: named propositions

Each proposition is a candidate for the Round 41 to 50 block, with its evidence and a test that could remove it.

- **P1. Structural abstention.** A composition can hold only if some component can produce HOLD or a disagreement can produce review. Evidence: E40-A 6.2. Removal test: a majority of non-abstaining witnesses that holds correctly at scale.
- **P2. Read, quote, then decide.** Models should locate and quote; a deterministic, span checked resolver should decide. Evidence: E40-A 6.3, T3 live 33 / 33 at 1 call. Removal test: a pool on which the resolver's fixed rules are systematically wrong where witnesses are right.
- **P3. Disagreement buys a hold, not a vote.** The second independent read should only be able to stop a release. Evidence: E40-A fault minimum 2; N01. Removal test: false hold rates on a pool with genuine extractor disagreement.
- **P4. Decorrelation is the safety budget.** A routed topology is as safe as its readers are decorrelated. Evidence: all 4 unsafe T6 states are identical misreadings. Removal test: measured joint misreading rates across witness pairs.
- **P5. Stages that can rewrite evidence add risk.** Evidence: X24. Removal test: validators restricted to flagging rather than rewriting.

11. Recommended Noetis structure (provisional)

1. Two independent extraction roles (different witnesses) return conditions, records, and verbatim quotes in a fixed schema; no verdict.
2. A deterministic gate checks every quote against the source, rejects any output off schema, and resolves in the fixed CEV-39 order with fail closed defaults.
3. Release only when both extractions resolve to the same determinate verdict; any disagreement, unverifiable quote, or HOLD goes to review.
4. Reviewers and critics may flag; they may not rewrite the evidence ledger.
5. Direct verdict witnesses, if used, act as an extra disagreement signal, not as a vote that can release.

This is the cheapest structure in this study with a two fault minimum, structural abstention, and no rewriting stage. It is recommended for testing, not adopted.

12. Rounds 41 to 50: proposed path toward Theory v1

This is a proposed research path. It is not a published commitment of the program.

Round	Proposed focus
41	Hard pool: items on which witnesses disagree, so topologies separate on accuracy
42	Measured decorrelation: joint misreading rates across witness pairs and prompts (P4)

Round	Proposed focus
43	Flag only validators versus rewriting validators (P5)
44	Abstention at scale with externally reviewed HOLD labels (P1, Round 39 handoff item 1)
45	Resolver failure: where fixed rules are wrong and witnesses are right (P2)
46	Review routing load: false hold rates and what a reviewer must see (P3)
47	Working memory and procedure reuse inside the routed topology
48	Transfer of the structure to a second domain outside contracts
49	Matched call baselines across all surviving propositions
50	Theory v1 draft: keep only propositions that survived Rounds 40 to 49

13. Limitations and threats to validity

- The pool is small (33 items) and mostly easy for these witnesses; live accuracy cannot rank topologies.
- New packets were authored by the research team. N02 has a plausible argument that the auditor's figure governs, which would make the truth FAIL rather than HOLD. No external reviewer audited expected verdicts before the run.
- Topologies share calls, so their errors are not independent samples.
- The K08 comparison with Round 39 changes more than one prompt element.
- Audit windows on real exhibits are excerpts; the X24 decider cited truncation.
- E40-A assumes a uniform state space and a fixed fault model; real error distributions are not uniform.
- Temperature 0 runs were made once per call.

14. Next check

Build the Round 41 hard pool before any further topology claim: items selected because the four direct witnesses disagree on them, with expected verdicts reviewed outside the authoring process, and with validators split into flag only and rewriting arms.

15. References

- [S09] Xuezhi Wang et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171, 2022.

- [S17] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325, 2023.
- [S24] Bingbing Wen, Jihan Yao, Shangbin Feng, et al. Know Your Limits: A Survey of Abstention in Large Language Models. arXiv:2407.18418, 2024.
- [S26] Pat Verga, Sebastian Hofstatter, et al. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. arXiv:2404.18796, 2024.
- [S28] A. P. Dawid, A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Applied Statistics 28(1), 1979. doi:10.2307/2346806.
- [S30] U.S. Securities and Exchange Commission. EDGAR Full-Text Search and Archives. <https://efits.sec.gov/LATEST/search-index> and <https://www.sec.gov/Archives/edgar/data/>.
- [S31] Geoffrey Irving, Paul Christiano, Dario Amodei. AI Safety via Debate. arXiv:1805.00899, 2018.
- [S32] Aman Madaan et al. Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651, 2023.
- [S33] Jie Huang et al. Large Language Models Cannot Self-Correct Reasoning Yet. arXiv:2310.01798, 2023.
- [S34] Lingjiao Chen et al. Are More LLM Calls All You Need? Towards Scaling Laws of Compound Inference Systems. arXiv:2403.02419, 2024.
- [S35] Krishna K. Ladha. The Condorcet Jury Theorem, Free Speech, and Correlated Votes. American Journal of Political Science 36(3), 1992.