Evidence Conflicts Before Verdicts

A bounded case study of evidence checks preceding collusion assessments

Flamur Koca | QENDROVA, Kiel, Germany

With Apart Research

ABSTRACT

Can a fixed evidence-to-verdict contract preserve distinctions between support, refutation, uncertainty and scope failure when manually mapped to new cases? During the 23-25 October sprint, we completed two offline exercises in two synthetic contexts. The 23 October restart evaluated six structured views; four admitted mappings produced 16 reference-matching labels and three open axes with opposite-valued compatible witnesses. Two views were manually stopped. A 24 October exercise evaluated six new prose cases prepared on 23 October; four admitted mappings produced 16 matching labels and five open axes, alongside two manual stops. These are correlated judgments, not independent trials. The first technical abort is retained. Pre-sprint model diagnostics and output-policy comparisons motivate, but do not constitute, these sprint results. All mapping and review were AI-assisted; the contract, source authority and coverage are stipulated. There is no independent holdout or human validation, no measured collusion outcome, and no established error reduction or net safety benefit. Sanitized prose-case evidence is available as L1; successful restart records remain local.

1. INTRODUCTION AND CONTRIBUTION

An audit can confuse evidence consistency with answer correctness or infer coordination from an authorization violation alone. We study the narrower evidence-to-verdict boundary, not a collusion detector. The new contribution is a bounded application of a preexisting contract to two case sets, preserving substantive uncertainty separately from manual scope stops.

Preexisting (19-22 October): G6 model responses, S1 and V8 checks, archived output-policy comparisons, GEGENWELT-3, and the fixed contract/core. Sections 2-4 and Appendices A-G report this foundation.

Sprint work: six structured views, references, failed first attempt and reviewed restart on 23 October (Section 5, H-I); six new prose cases and references prepared on 23 October, executed and reviewed on 24 October (Section 6, J-K). Section 7 provides a retrospective synthesis of these completed results; it is not another experiment.

2. RELATED WORK

WikiContradict [1] studies conflicting knowledge; our synthetic equal-authority records are narrower. Selective classification [2] relates abstention to coverage; our retrospective counts use all planned answers, not accepted answers, and are not selective-risk estimates. Secret Collusion among AI Agents [3] studies hidden coordination; we observe no experimental agent-coordination process, hidden communication, or collusive incentives. This focused comparison is neither replication nor an exhaustive literature review. Appendix A retains the detailed distinctions.

3. PREEXISTING METHODS

G6 contains six archived Qwen3-4B-Instruct-2507 Q4_K_M responses: three synthetic worlds under tasks F and I, once per pair without repeated trials or replacements. F requests validity and plan acceptance, execution, rule violation, and plan binding. I requests only validity, error code, and rationale. Within pairs, facts and system rules match; task and output scope differ.

Each world has two equal-authority records of one action, not two actions or a correction sequence. K reports EXECUTED/FAILED and requires INVALID_INPUT/AUTHORITATIVE_CONFLICT without substantive verdicts. P reports EXECUTED/EXECUTED and N FAILED/FAILED; both require VALID/null. Source authority, action identity, complete channels, empty communication/authorization logs, and null planReference are synthetic assumptions, not verified real-world properties.

S1 was developed after observing the errors. Its narrow structured contract blocks K, passes one executed action with both sources for P, and retains N's negative evidence without an executed action. Separate V8 checks support execution and the authorization violation for P, and refute both for N. There are 81 passing offline checks, not 81 independent model cases or a blind evaluation. S1 does not establish source truth or correctness of arbitrary text and plan judgments.

Archive arm A releases every original response; B suppresses it when the archived S1 status is BLOCKED_CONFLICT, otherwise leaving it unchanged. No new response, repair, or V8 replacement is involved. READY_FOR_V8 is not answer correctness.

A separate offline demonstration uses six constructed candidates on the three known worlds, four sharing K. H suppresses blocked candidates; R exposes them unchanged only in an unreviewed, blocked channel without action authority. Unblocked controls, sources, and status notices match between arms. Statuses are stipulated, not new S1 runs. Twelve routing rows use predetermined AI references known to reviewers; the evaluator consumes classes, not semantic judgments. Counts concern sentence units, not independent facts. Full settings, check composition, references, and source bindings remain in Appendices B-E.

4. PRE-SPRINT RESULTS

Both K responses incorrectly return VALID/null. F/K mentions both outcomes but treats one record as a duplicate and later uses it for a rule verdict; I/K gives no rationale. P/N validity is correct. F/P has a wrong definitive rule verdict, wrong open plan judgments, and a format error. F/N correctly refutes execution and violation but has wrong open plan judgments and describes two records as two actions. Validity/code accuracy is F 2/3 and I 2/3; format compliance F 2/3 and I 3/3. Unasked I judgments are not accuracy gains. Original quotations and all six response IDs are retained in Appendix C.

Table 1. Archived response selection. A releases all answers; B suppresses conflict cases. Each task retains a denominator of three planned responses.

Measure	F: A -> B	I: A -> B
Response coverage	3/3 -> 2/3	3/3 -> 2/3
Responses with a definite primary error	2/3 -> 1/3	1/3 -> 0/3
Fully correct released responses	0/3 -> 0/3	2/3 -> 2/3
Released format errors	1/3 -> 1/3	0/3 -> 0/3

A primary error is at least one wrong definitive judgment, counted once per answer. Incorrect OFFEN can prevent full correctness without meeting that definition. Content/format counts overlap; F/I are not pooled. Both wrong K answers are withheld, while F/P and F/N errors remain. The archive contains no correct conflict rejection; B would suppress one too. Its loss rate is unmeasured.

Table 2. Constructed candidates by output channel. H suppresses blocked candidates; R exposes them only in a blocked review channel. Counts are sentence units, not independent cases.

Sentence-unit class	H normal	H review	R normal	R review
Correct conflict hints	0	0	0	2
False units	2	0	2	4
Unsupported claims	0	0	0	1
Other correct units	2	0	2	1

Both correct hints concern the same conflict. R02 combines one with two false sentences; it is not correct overall. The nonfactual "Bitte prüfen" is excluded. Sources and status remain available in both arms. No information rescue, human benefit, net-benefit score, or winning arm is established. Appendix D retains the per-arm totals and interpretation limits.

5. SPRINT-PERIOD TRANSFER RESTART

On 23 October, a six-view exercise used two synthetic contexts: publication approval and data export. Each has a full/partial-view pair; two additional views probe a rule-scope boundary and unresolved authority. These are six views, not six independent incidents. Their language is strongly schema-shaped, with explicit channels, completeness, event order, and plan references. They are not natural incident narratives.

The first attempt stopped technically before any core execution. A separately authorized restart reused these same cases, references, fixed answer contract, and preexisting core. Six fresh AI mapping roles received their own view after an acknowledgement step, without inherited chat history or reference answers according to the process record. Initial mappings were captured, then four admitted inputs were sealed and executed once. Reference comparison followed execution. There were four core calls, no retries, and no research-model/backend calls; AI mapping and review were used.

View	Manual scope	Plan	Execution	Violation	Plan link
V01	IN_SCOPE	BELEGT	BELEGT	WIDERLEGT	BELEGT
V02	IN_SCOPE	OFFEN	BELEGT	WIDERLEGT	OFFEN
V03	IN_SCOPE	BELEGT	BELEGT	BELEGT	BELEGT
V04	IN_SCOPE	BELEGT	BELEGT	OFFEN	BELEGT
V05	OUT_OF_SCOPE	--	--	--	--
V06	UNCLEAR	--	--	--	--

BELEGT / WIDERLEGT / OFFEN mean supported / refuted / open within the fixed contract, not real-world truth. The four axes are PLAN_ACCEPTED, CLOSE_EXECUTED, RULE_VIOLATED, and PLAN_ACTION_LINKED. Dashes mean no core judgment, not OFFEN or a failed judgment.

All 16 labels match the prepared references. V02 has two open axes because the other participant's acceptance is unobserved; V04 has an open violation axis because the authorization channel is unobserved. Each open axis has compatible witnesses with opposite target values. Witness positions can differ from the reference example while remaining valid. V05 introduces withdrawal of plan acceptance, outside the contract. V06 leaves the competent authority unresolved. Those two stops were manual, not demonstrated automatic detection.

The old abort remains a separate negative process result: invalid V01 first mapping, lost pre-edit V03 raw bytes, and zero core executions. Its unexecuted outcomes are not converted to open judgments or a zero-percent score. Appendix H preserves that historical record; Appendix I records the restart and its limits.

6. LANGUAGE TRANSFER WITH NEW PROSE CASES

The separate 24 October exercise used six synthetic prose cases L01-L06, prepared on 23 October with separate reference answers and prespecified assessment criteria. It retained the same publication-approval and data-export contexts, fixed answer contract, and deterministic core. The cases use less schema-shaped wording, but still stipulate evidence authority, channel coverage, event order, and relevant identities. This is not multilingual transfer or unconstrained real-world text extraction.

Six fresh AI mapping roles received only their own case and neutral mapping materials after acknowledgement, without inherited chat history or supplied reference answers. First submissions were captured unchanged and sealed before execution and reference comparison. Four eligible inputs were evaluated once; two manual scope stops produced no core judgment. There were no retries or research-model/backend calls; AI mapping and review were used. Execution ended eight minutes after authorization; team review closed within nineteen minutes, below the two-hour cap. Appendix J records provenance and limits.

Case	Manual scope	Plan	Execution	Violation	Plan link
L01	IN_SCOPE	BELEGT	BELEGT	OFFEN	BELEGT
L02	UNCLEAR	--	--	--	--
L03	IN_SCOPE	OFFEN	WIDERLEGT	WIDERLEGT	WIDERLEGT
L04	OUT_OF_SCOPE	--	--	--	--
L05	IN_SCOPE	WIDERLEGT	BELEGT	WIDERLEGT	WIDERLEGT
L06	IN_SCOPE	OFFEN	BELEGT	OFFEN	OFFEN

Labels and dashes have the fixed-contract meanings defined in Section 5. All 16 evaluated labels matched the separate references. Review supported six scope classes and reasons, eight visible communication typings, and five open-axis witness pairs: L01 violation, L03 plan acceptance, and L06 plan acceptance, violation, and plan link. These are correlated judgments, not independent trials.

L02 preserves an unknown initial grant despite a complete empty register within the later window. L04 includes an effective grant revocation, outside the contract. L03 does not turn a failed close or successful audit read into target execution. L05 separates proposal, quotation, and late acceptance. L06 does not interpret a quotation without present acceptance as a claim of never accepting. Review found no material mapping error, omission, unsupported addition, or reference dispute. The two scope stops are not extra label successes. This supports the tested manual mapping pipeline, not an automated semantic safeguard.

7. WHAT THE SPRINT EXERCISES ADD

This retrospective synthesis interprets the completed 23-24 October results; it is not a new run or a controlled comparison of wording. The contribution is narrower than collusion detection: a documented manual mapping pipeline preserved selected evidence-to-claim distinctions under a fixed contract, including cases where no core judgment was appropriate.

Stage	Observed result	What a reviewer can inspect
Pre-sprint, 19-22 Oct	G6/S1, routing and three counterworld pairs; not new sprint findings.	R2: model archive/routing. R1: counterworlds and controls.
23 Oct first attempt	Technical abort; zero core executions. Not an OFFEN verdict or 0% accuracy.	A1: sanitized failure account; not raw logs.
23 Oct restart	Six structured views: four admitted mappings, 16 matching labels, three open axes; two manual stops.	Section 5 and Appendix I. Original restart records remain local, outside the four public releases.
24 Oct prose exercise	Six new prose cases: four admitted mappings, 16 matching labels, five open axes; two manual stops.	L1: cases, mappings, references, core results and assessment. Operational captures remain withheld.

A concrete contrast: quotation is not acceptance

L05 has complete communication coverage. The early text is a proposal and a quoted acceptance, not an actual acceptance. Actual acceptances occur at positions 23 and 25, after execution at 16. Prior plan acceptance is therefore WIDERLEGT (refuted) within the contract. L06 instead has partial communication coverage: a quotation at 10 is not acceptance, but does not exclude a separate unseen earlier acceptance before execution at 20. Prior plan acceptance remains OFFEN (open). This distinction is supported by L1 CASES.json, REFERENCES.json, MAPPINGS.json and CORE-OUTPUTS.json (L05/L06); it is not a newly observed model error.

A boundary is not an uncertain verdict

L02 lacks a required initial authorization status; its complete empty register covers the later window only. Manual screening returns UNCLEAR rather than guessing a core verdict. L04 permits grant revocation, outside the fixed contract, and is stopped as OUT_OF_SCOPE. These stops are distinct from the open axes admitted in L01/L03/L06 and are not extra label successes.

Interpretation: correct manual mapping and matched outputs on these cases support a limited contract application. They do not establish autonomous semantic detection, transfer to new domains, reduced errors against a baseline, or collusion. Wording and content changed together, and the two sets of 16 labels must not be pooled as 32 independent trials.

8. DISCUSSION AND LIMITATIONS

Mentioning a conflict does not ensure its use in a verdict; input validity does not certify answer correctness. Suppression trades coverage for fewer visible errors, while review exposes mixed-quality text. Neither establishes a better model or complete audit system. The model's internal causal mechanism remains unknown.

The 23 October restart reused cases from the abort in two structured contexts. The 24 October exercise adds new prose cases in those same contexts, not an independent external holdout. Wording and content changed together; a causal benefit of wording cannot be isolated. Neither establishes transfer to real incidents or new rules. Each set of 16 labels is correlated across four mappings, not a pooled accuracy estimate. AI-authored cases, references, mappings, and reviews are not independent human validation.

There is no without-procedure baseline, estimate of error reduction, net-benefit measurement, or measured collusion outcome. Authority, completeness, and identity are stipulated. Hashes bind bytes, not source truth or unobserved reference access. Role separation is documented, not cryptographic proof of blinding.

Semantic agreement is not complete output-format compliance. Raw core results lack some required per-axis evidence, assumptions, reasons, missing-information, and alternative-description fields. Visible event lists are not minimal axis-specific evidence sets. Reviewer explanations are post-hoc derivations, not core output. Manual scope screening is not automatic semantic detection. False accusations, misplaced confidence, lost warnings, review-channel influence, and evasion remain risks; no live deployment or automatic sanctions are justified.

The separate twelve-case, 24-response model follow-up remains unexecuted; these offline exercises do not repair its backend-binding limitation. Historical diagnostic details remain in Appendix E.

9. CONCLUSION

Pre-sprint diagnostics show conflict-validity failures and output-policy trade-offs. The 21 October counterworld check supported three known partial-view judgments and rejected three defective controls. The 23 October restart and 24 October language transfer each yielded four matching in-scope results and two supported manual stops, with three and five open axes respectively. The original abort remains a negative process result. None demonstrates collusion detection, general transfer, or net safety benefit.

CODE AND DATA

R2 (sanitized reading evidence):
https://drive.google.com/file/d/1QASGzR7xfGY8Zs4jHwvVUNWobBWgXE7G/view

R1 (counterworld evidence): https://drive.google.com/file/d/1CNAhJkuzg3TBd7iFgP-1PXQbXsfj0IWA/view

A1 (historical abort): https://drive.google.com/file/d/1toeXBU4BuJ4qsaFEZymxPj5UKO9dLUeb/view

L1 (L01-L06 language-transfer reading evidence):
https://drive.google.com/file/d/1WLJZ42azGF8_6EXYz81FRuhTG5TKOwuf/view

R1/R2/A1 retain their earlier scope; L1 covers only L01-L06. These releases are not complete executable reproductions. Operational records remain withheld. Appendix K explains L1's historical report binding. Prior reports and original evidence are retained unchanged; this final report distinguishes their dated findings from completed sprint work. The original submission was received on 24 October 2026.

References

[1] Hou et al. (2024). WikiContradict: A Benchmark for Evaluating LLMs on Real-World Knowledge Conflicts from Wikipedia. https://arxiv.org/html/2406.13805v1

[2] El-Yaniv, R., and Wiener, Y. (2010). On the Foundations of Noise-free Selective Classification. Journal of Machine Learning Research, 11, 1605-1641. https://jmlr.org/papers/volume11/el-yaniv10a/el-yaniv10a.pdf

[3] Motwani et al. (2025; first posted 2024). Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. https://arxiv.org/html/2402.07510v5

Appendix

A Extended related work

WikiContradict [1] studies LLM handling of conflicting Wikipedia knowledge. Our case is narrower: synthetic records with explicitly equal authority and fixed validity rules. It is not a replication of that benchmark.

Selective classification [2] formalizes abstention and coverage. Our suppression policy is fixed and retrospective, not learned. Our error counts use all planned answers as the denominator, not only accepted answers; they are therefore not estimates of selective risk or statistical guarantees.

Secret Collusion among AI Agents [3] studies multi-agent deception via steganography. We do not observe a multi-agent coordination process, hidden communication, or collusive incentives. The connection is prospective: unreliable evidence processing can undermine a later audit. This is a focused comparison, not an exhaustive literature review or a claim of novel collusion detection.

B Detailed methods and settings

B.1 Archived model diagnosis (G6)

Qwen3-4B-Instruct-2507, Q4_K_M GGUF, produced one response for each of three synthetic worlds under two tasks: F requested validity and four substantive judgments (plan acceptance, execution, rule violation, plan binding); I requested only validity, error code, and rationale. Facts and system rules were identical within each pair, but task and output scope differed. There were six responses, no repeated trials, and no replacement responses.

Configured limits were an 8,192-token context, at most 5,000 input and 1,200 output tokens, 150 seconds per call, and no retry. Configuration specified temperature 0.2, top_p 1, seed 20261019, and zero presence/frequency penalties. These settings do not establish deterministic backend execution or robustness across seeds.

Each world concerns action act-42, actor B, task T, action position 12. Records e-k1 and e-k2 were registered at positions 14 and 19. Both are authoritative; neither supersedes the other. Communication, event, and authorization channels are stipulated complete. Communication and authorization logs are empty, and planReference is null. These are synthetic assumptions, not verified real-world properties. World K reports EXECUTED/FAILED; P reports EXECUTED/EXECUTED; N reports FAILED/FAILED. K requires INVALID_INPUT/AUTHORITATIVE_CONFLICT with no substantive verdicts. P and N require VALID/null.

B.2 Structured check S1 and archived output selection

S1 was developed after observing the model errors. Its narrow contract handles structured statements, specified sources, roles and windows, with no grant entries and fixed statement forms. It preserves source-to-action attribution. K yields BLOCKED_CONFLICT and no downstream V8 input. P yields one executed closure referencing both records. N yields no executed action while retaining negative evidence. Separate deterministic V8 checks support execution and the violation CLOSE without a strictly earlier R-GRANT for P, and refute both for N.

Archived offline checks comprise 21 unit checks, 45 separately developed counterchecks, and 15 V8 regressions, all passing. Counterchecks were created before reading the new implementation; reference values were not adjusted to obtain passes. Contracts were known. These are not 81 independent model cases or a blind evaluation. S1 does not validate arbitrary text, source truth, all plan judgments, or the real-world correctness of an action identifier.

In a separate archive comparison, arm A releases every original G6 response. Arm B withholds the entire response when its archived S1 status is BLOCKED_CONFLICT; otherwise it releases the original unchanged. No new response, repair, or V8 replacement was produced. READY_FOR_V8 means eligible for a handoff, not that the model response is correct.

B.3 Constructed review-channel demonstration

Six hand-written candidates use the three known worlds; four share K. These are not additional model outputs or six independent worlds. H suppresses a blocked candidate entirely. R exposes that same candidate only in a separate review channel marked unreviewed and blocked, without permission to act. Both arms release the same candidates normally for unblocked controls. Sources, status notices, and candidate content are held fixed within each pair.

Input statuses are stipulated contract expectations, not new S1 executions. Twelve routing rows were evaluated offline against separate, predetermined AI-generated references after complete channel assignment. Reviewers knew the references. The evaluator consumed reference classes rather than recognizing semantic correctness. Counts use fixed sentence units, not independent atomic facts. Unsupported claims remain separate from demonstrably false statements.

C Original cases and archive interpretation

C.1 Three original cases

K: Both F/K (G6:1) and I/K (G6:4) incorrectly return VALID/null. F/K says: “Der Beleg e-k1 hat outcome 'EXECUTED', aber wird als doppelter Eintrag behandelt.” It later uses e-k1 for the rule-violation judgment. Both outcomes are mentioned, but the stipulated conflict does not reach the validity verdict. I/K provides no rationale, so we cannot infer whether it internally recognized the conflict.

P: F/P (G6:5) and I/P (G6:2) return the correct VALID/null. F/P correctly labels execution but labels RULE_VIOLATED as WIDERLEGT (refuted), although its rationale says “wird die Regel verletzt.” Plan acceptance and plan binding are incorrectly OFFEN (open). Nonempty missingInformation beside a definitive label also violates the output contract. Content and format failures coexist.

N: F/N (G6:3) and I/N (G6:6) return the correct VALID/null. F/N correctly refutes execution and rule violation but incorrectly leaves the plan judgments open. Its wording, “Beide CLOSE-Aktionen für T im X-Log”, describes two records of one action as two actions. Plan acceptance is refuted by the stipulated complete, empty communication log, not by planReference:null alone.

Validity/code accuracy is 2/3 in each task. Format compliance is F 2/3, I 3/3. The narrower task did not fix this validity failure. Unasked substantive judgments in I are not accuracy gains. The internal causal mechanism is unknown.

C.2 Post-hoc selection, without repair

Denominator: three originally planned responses per task, including withheld ones.

A primary error counts at most once per response and requires a demonstrably wrong definitive judgment. An incorrect OFFEN can prevent full correctness without meeting that narrower definition. Content and format counts overlap and must not be added. F and I are not pooled into one performance score.

Both wrong K responses are withheld. F/P retains content and format errors; F/N retains incorrect plan judgments. Correct I/P and I/N remain available. No fully correct response was withheld in this archive, but it contained no correct conflict rejection. B would suppress such a rejection too; that is a contract consequence, not a measured frequency or a cost-free safety result.

D Routing and separate counterworld certificates

Per-arm candidate totals are two correct conflict-hint occurrences, six false units, one unsupported claim, and three other correct units. One nonfactual “Bitte prüfen” request is excluded. Both conflict hints concern the same known conflict. Mixed candidate R02 contains a correct hint and two false sentences; it is not a correct answer overall.

R exposes correct candidate statements together with misleading content. Both arms retain the original sources and common conflict-status notices. We did not establish information rescue for a reader, improved comprehension, timely use, lower workload, or lower harm. This is a completed demonstration of specified routing, not a human-effectiveness experiment. There is no net-benefit score or winning arm.

D.1 Separate counterworld certificates: pre-sprint work on 21 October

GEGENWELT-3 is separate from the G6 model archive and constructed routing demonstration. Three known synthetic views were selected from the existing reference package: F11/v-c5a830 (PLAN_ACCEPTED), F03/v-41b7da (CLOSE_EXECUTED), and F08/v-ab0296 (RULE_VIOLATED). Existing true/false completion recipes were used, not newly sampled or unknown cases. Each reconstructed world contains exactly the visible events plus its declared additions, closing K/X/R within positions 0..30. The original PARTIAL/NONE coverage remains unchanged. Closing a hypothetical world does not establish that the original view was complete.

The checker first requires unchanged visible records, no additions to COMPLETE channels, unique event identities and positions, and at most one executed CLOSE(T). It then derives the selected target value before comparing separate references. F11 uses exact, disclosed manual classifications of acceptance utterances; F03 and F08 use structured action/authorization fields. A failed CLOSE or successful AUDIT_READ does not count as executed CLOSE. Authorization requires GRANT(T) by R strictly before CLOSE(T). A violation is not itself an invalid world. Reference values were known to the team; this was not blinded.

One authorized offline run processed W1,W2,N1,W3,W4,N2,W5,W6,N3 on 21 October 2026, without retry, code repair, replacement cases, or model/backend calls. F11: W1=true with the added B-to-A acceptance e-21a; W2=false without it. F03: W3=true with added alt-close-20 (CLOSE T EXECUTED); W4=false without it. F08: W5=true without a valid earlier grant; W6=false with alt-grant-3 at position 3 before the visible CLOSE at position 23. All six worlds were compatible with their respective original views. All three pairs changed the target truth value as specified; no reference mismatch occurred.

The three controls are additional defective copies, not extra independent cases. N1 changes an added event to an occupied position: INVALID_WORLD / POSITION_COLLISION. N2 adds a K event despite original K=COMPLETE: INCOMPATIBLE_VIEW / COMPLETE_CHANNEL_ADDITION. N3 changes a visible event's task from T to U only in the world copy: INCOMPATIBLE_VIEW / VISIBLE_RECORD_CHANGED. All were rejected for the specified reason and returned targetValue=null. No missing objects or technical stop occurred.

Two internal AI reviews inspected the saved results and source bindings without rerunning the checker. For these three selected target statements, opposite-valued compatible completions establish underdetermination within the declared contract. This is not a result about all possible inputs or a measurement of model improvement, collusion, intent, guilt, human benefit, or live safety. Manual language typing, stipulated authority/completeness, known cases and the absence of blind or independent human validation limit the claim. The controls cover only their three defect types. Consistent checkpoints are not an externally anchored tamper-proof audit trail. Both implementation and execution were brought forward to 21 October by explicit user authorization: both are pre-sprint work, not sprint results.

E. Limitations and unexecuted follow-up

The archive illustrates why mentioning a conflict is not the same as applying it to the verdict, and why a valid input status does not certify an answer. A narrow source-attribution check can support a conservative output policy, but this retrospective result does not show a generally better model or a complete audit system. Suppression exchanges coverage for fewer visible errors; a separate review channel exchanges access for exposure to mixed-quality text.

Limitations

The G6 source worlds are small, synthetic, and known; S1 was developed after observing the G6 errors. The six routing candidates were constructed later. There is no independent unknown holdout, independent human reference assessment, external replication, statistical estimate, or measured collusion outcome. Source authority, completeness, and action identity are assumed. File hashes bind artifacts, not the truth of their contents. The exact error mechanism cannot be inferred from these outputs.

Future Work

Future work already prepared locally includes twelve fixed inputs and two repetitions (24 planned responses), covering four conflict/consistency pairs and four separately scored boundary controls. It is not an unknown holdout. No responses were collected: the required live instance/engine/process binding was not established. A load-observe-unload diagnostic produced CLI exit 0xC0000005 and no runtime stdout bytes. The cause is unresolved. A sanitized issue is recorded at https://github.com/lmstudio-ai/lms/issues/659. Local stopping or discarding late output would not prove backend cancellation. This follow-up is not an executed experiment or a condition for honestly reporting the existing archive. No execution approval is implied here.

F. Code and data availability

The sanitized archive QENDROVA-Apart-Belege-20261021-R2.zip is available at: https://drive.google.com/file/d/1QASGzR7xfGY8Zs4jHwvVUNWobBWgXE7G/view Its root directory is BELEGPAKET-BEREINIGT-20261021-R2/. START-HERE.txt explains the reading order; MANIFEST.json records file hashes, source provenance, declared redactions, and the archived pre-link report hash. LESERBELEGE-20261021/ contains six full message sets and original responses, separate references, contracts, configuration, S1 records and archive selection. PRUEFAUSGABE-6-20261021/ separately contains constructed candidates, references, archived routing and counts. No evaluator or execution runner is included. diagnostik/ contains a declared derivative of the failed diagnostic report.

This is not a complete executable reproduction archive: S1 implementation and test sources are excluded. Link access to this unchanged R2 package was authorized before the original report submission on 24 October. Private conversations, raw system logs, memory dumps and unapproved internal notes are excluded. Appendix G maps claims to packaged reading evidence, not to inaccessible private log locations.

Access check, 21 October 2026: the file page returned HTTP 200 and the ZIP was downloaded without Google authentication or browser cookies. Its 68,859 bytes matched the unchanged local R2 archive. Link holders can download and forward the package; access is not restricted to named reviewers. R2 ZIP SHA-256: 2c0efe42a7461dfd97abf99ca6441ee49c433a5d710a4bf75317501313777b99

Version binding: the unchanged R2 manifest identifies the archived report QENDROVA-Apart-Report-20261021-4p-Belege.pdf, SHA-256: 00444359e68ed6594164fb7197eda985fdeaa3b5aff71c45d38c1f5d628c2c51 It does not bind this

later revision. The earlier link-and-abstract revision only changed presentation and evidence availability. The 21 October counterworld revision added the separately identified pre-sprint GEGENWELT-3 finding in Appendix D. The 22 October link revision changed availability, not the pre-sprint results. R2's local-only and pending-release notices describe its earlier packaging state; the subsequent link-release authorization is recorded here explicitly. The ZIP has not been repacked. This report identifies that exact ZIP by hash.

The separate sanitized counterworld archive QENDROVA-Apart-Gegenwelt-Belege-20261021-R1.zip is available at: https://drive.google.com/file/d/1CNAhJkuzg3TBd7iFgP-1PXQbXsfj0IWA/view It supplements, and does not replace, R2. Its root is BELEGPAKET-GEGENWELT-20261021-R1/. The 13 files include START-HERE.txt, BERICHT-ZUORDNUNG.json, MANIFEST.json, selected views and completion recipes, manual K typing, controls, separate references, all nine sanitized result objects, checkpoint projections, a source map, methods extract and core.mjs. No runner, private authorization text, raw logs or unrelated source collections are included. This is a reading and inspection package, not a complete executable reproduction archive. No independent reproduction is claimed.

Access check, 22 October 2026 (Europe/Berlin): Google Drive showed the QENDROVA account as owner and Anyone with the link / Viewer. An unauthenticated download without browser cookies or Google credentials returned HTTP 200 and 29,271 bytes, matching the original ZIP. Its SHA-256 is: 454ee3f9cecac77739ed48b9630432146abc6b1d57d39d4491861d4e69cb2c60 The manifest SHA-256 is: 1a57ab0e9d6999ca94f8f7bb648b6762bbefab514894b1443a6bda7584d39930 Link holders can download and redistribute it; access is not limited to named reviewers. Owner identity may be visible. Link release was authorized on 22 October; the original report was subsequently submitted on 24 October.

Version binding: the counterworld R1 manifest binds the archived report QENDROVA-Apart-Report-20261021-4p-Counterworld.pdf, SHA-256: 04e39475b8b90fd6af131990e09a6703fb5c0b92852b9941bdd8cb02d0ad01f4 It does not bind this final report. Local-only and pending-release notices inside the unchanged ZIP describe its earlier packaging state; the subsequent authorized link release on 22 October is explicitly recorded here. The ZIP was not repacked. Its source map retains hashes of original artifacts; hashes bind bytes, not source truth or an externally anchored audit trail.

G. Dual-Use / Ethical Considerations and chronology

Unsupported collusion labels may falsely accuse people or systems. A formally valid record may create unwarranted confidence. Suppression may hide correct warnings; review labels do not demonstrate protection from influence. Known checking rules may invite evasion. Neither live deployment nor automatic sanctions are justified by these findings. Privacy review and explicit release approval remain necessary for any external evidence package.

19 October 2026: G6 diagnosis, S1 implementation and offline checks. 20 October: retrospective archive comparison and follow-up preparation. 20-21 October: technical preparation/diagnosis without follow-up model answers. 21 October: reader package, constructed routing demonstration, report assembly, this English text transfer, and the separate GEGENWELT-3 checker construction and single offline run. Six counterworlds plus three defective control copies were processed; no model/backend calls were made for GEGENWELT-3.

22 October: the sanitized counterworld R1 package was released separately through a QENDROVA-owned Viewer link and its anonymous download was verified. That dated link revision updated availability and version bindings only. R2 and the archived reports remain unchanged. No new research run occurred.

The activities above are pre-sprint work. During the sprint, the first six-view attempt aborted on 23 October; the separately authorized restart completed later that day. New prose cases and references were prepared on 23 October and evaluated and reviewed on 24 October. Sections 5-7 and Appendices H-K distinguish these contributions and their limits. The original report was submitted on 24 October. This revision corrects report status and clarifies the existing results; it adds no research run. Registration and survey completion are administrative steps, not research findings.

Local provenance map (relative to BELEGPAKET-BEREINIGT-20261021-R2/; these are packaged file paths, not public download links):

- Entry and provenance: START-HERE.txt and MANIFEST.json. L below means LESERBELEGE-20261021/; R means PRUEFAUSGABE-6-20261021/.

- G6 inputs: L/g6/eingaben.json. Original response strings G6:1-G6:6: L/g6-originalantworten.json. References: L/g6/sollreferenzen.json.

- Settings and ratings: L/g6/laufkonfiguration-lesefassung.json and L/g6/bewertung-auszug.txt. Requested schemas/template: L/g6/schema-voll.json, L/g6/schema-isoliert.json and L/g6/prompt-template.json. Contracts: L/vertraege/.

- S1: L/s1/ contains K.original.json, P.original.json, N.original.json and the corresponding *.ergebnis-lesefassung.json; archived check records: archivierte-pruefzusammenfassung.json and archivierte-softwarepruefung.txt.

- Archive selection (Table 1): L/archivvergleich-auszug.json.

- Constructed routing (Table 2): R/FAELLE.json, R/REFERENZEN.json, R/ROUTING.json and R/ERGEBNIS.json. These are not new model responses.

- Failed follow-up diagnostic: diagnostik/bericht-bereinigt.json. Client record only; no independent backend or cancellation proof.

Counterworld provenance map (relative to BELEGPAKET-GEGENWELT-20261021-R1/): eingaben/ contains the three views, six recipes, manual K typing and three controls; referenzen/sollwerte.json contains separate expectations. ergebnisse/lauf-bereinigt.json retains all nine outcomes and three case results; ergebnisse/ablauf-auszug.json contains checkpoint projections, not raw logs. quellen/source-map.json records original-source hashes and redactions. vertrag/methoden-auszug.txt and code/core.mjs expose the declared method and pure checking logic, without the original runner. BERICHT-ZUORDNUNG.json maps the results to report claims; MANIFEST.json binds the packaged files and the archived pre-link counterworld report. This does not certify source truth.

LLM Usage Statement

AI assistants supported study preparation, software work, internal reference construction, counterchecks, interpretation, drafting, and this English transfer. The evaluated G6 responses are distinct from these assistance activities. Internal AI reviews are not independent human validation. The author remains responsible for checking all claims, sources, translations, attribution, and release permissions. The original report was submitted with author authorization on 24 October; this corrected final report retains the same responsibility and disclosure boundaries.

H. TECHNICAL ABORT ON 23 OCTOBER 2026

Historical negative process result, preserved separately from the successful restart

Scope and chronology

This appendix preserves the first sprint-period attempt, which stopped before core execution. It is separate from the successful same-day restart in Appendix I and the 24 October prose-case exercise in Appendix J. Earlier report files and original run evidence remain unchanged; the current report presents the completed sequence.

Attempt and outcome

On 23 October, a separate six-view transfer comparison began at 16:26:01 UTC and was documented as stopped at 16:43:26 UTC, within its two-hour cap. Six views and separate references were prepared, but the attempt stopped before core execution. There were zero core calls and zero research-model/backend calls. All six core outcomes remain NOT_EXECUTED, with null labels and scores: neither an OFFEN (open substantive verdict) nor 0% accuracy. This aborted attempt yielded no semantic performance result. It remains a negative process result and is not relabeled as the later successful restart.

Documented process failures

Serialization instructions were incomplete, including omission of required enum values. Separately, incorrect PowerShell escaping supplied by the coordinating assistant produced literal backslashes and invalid initial JSON files for V01 and V03. V03 was mechanically edited before its original bytes were separately preserved. An unchanged V03 first-run artifact, and semantic equivalence between its initial and edited versions, therefore cannot be independently established. This limitation concerns V03 in this attempt, not the earlier archived cases. No repaired first run or automatic retry is claimed.

Interpretation and limits

This attempt is distinct from the twelve-input, 24-response model follow-up described in the historical report; it neither executes nor replaces that follow-up. It does not overturn the earlier bounded findings, validate transfer or collusion detection, or establish a safety benefit. The A1 materials document sprint-period preparation and this technical abort, not the later completed semantic evaluations. This account is based on local process records and internal AI counterchecking, not independent human validation or a new experimental run.

Evidence and release boundary

Local sources remain ABBRUCH-20261023.txt and STATUS.json. A separately authorized sanitized derivative, QENDROVA-Apart-Abort-Record-20261023-A1.txt, is available through the Viewer link below. It supports this abort account, not a completed semantic evaluation. It does not disclose the original logs, cases, reference answers, mappings or software. A1 is separate from R1/R2; their archives, links, permissions and historical report bindings remain unchanged. https://drive.google.com/file/d/1toeXBU4BuJ4qsaFEZymxPj5UKO9dLUeb/view A1: 4,791 bytes. SHA-256: ed4f9fa41167c2836d0ee56231e4da785b8b99a3f0d1470afa7b4900b7cbd5f9. On 23 October 2026, an unauthenticated download without browser cookies or supplied credentials returned HTTP 200 and matched the authorized A1 bytes and hash. Internal AI counterchecking confirmed the saved download's byte equality, not an independent network fetch. This is a point-in-time access check. Link holders may download and redistribute the file; owner identity may be visible. Version binding: unchanged A1 identifies the preceding QENDROVA-Apart-Report-20261023-Abort-Addendum.pdf, SHA-256 b117cf864c15f68c5c4d711e00d84c77a855868612e3db43dfbb6326412e7fb4, not this final report. The 23 October A1 link update changed evidence availability, not findings. A1 remains the released abort record; it does not establish the results of the successful restart or the prose-case exercise. Their evidence boundaries are stated separately in Appendices I-K.

I. 23 OCTOBER RESTART: PROVENANCE AND LIMITS

The archived reports and all extant original run files remain retained. Appendices A-G describe the preexisting foundation and its evidence; Appendix H preserves the first technical abort. This appendix reports the separately authorized successful restart on 23 October, not a correction or deletion of that failed attempt.

I.1 Chronology and execution boundary

Preexisting work includes the fixed answer contract and deterministic core dated 22 October, as well as the earlier G6, S1, routing, and counterworld work. The six views, prepared references, first attempt, corrected restart procedure, mappings, execution, and reviews are dated 23 October. The restart reused cases from the failed attempt: it is not an untouched test set or independent replication.

The first attempt ran from 16:26:01 to 16:43:26 UTC and stopped before core execution. Incomplete enum instructions and controller-side PowerShell escaping contributed to the failure. The invalid first V01 mapping was retained; V03 was mechanically edited before original bytes were preserved, so the unchanged first-run bytes and their semantic equivalence cannot be independently established. V02 was not written, and V04-V06 were not started. These failures remain reported as failures.

The separately authorized restart began at 19:56:45 UTC with a two-hour limit. Its four logged core calls finished at 20:06:35.972 UTC. The controller recorded first reference opening at 20:06:40 UTC; all three team reviews were received by 20:10:55 UTC. There were no retries or research-model/backend calls. Mapping and review were performed by AI roles; this was not an AI-free study or human validation.

I.2 What was checked

Captured initial mappings, receipts, source bindings, six acknowledgements, admitted inputs, and execution records were compared offline without repeating the core calls. Semantic review checked all 16 labels, the three open axes and their opposing witnesses, and the two manual scope stops. The reviewers reported no new material mapping or process error. That finding is limited to these records, not a guarantee that no error exists.

V02's unobserved acceptance supports both absence and a compatible earlier acceptance; both its plan-acceptance and plan-link judgments remain open. V04's missing authorization channel supports worlds with and without a timely grant. V05 concerns plan-acceptance withdrawal, not grant revocation or expiry. V06 cannot assign authority without the missing protection class. No verdict is fabricated for either stopped view.

I.3 Audit and disclosure boundaries

Capture hashes bind saved bytes. They do not prove first thoughts, source truth, or absence of prior exposure. Some mapping/admission timestamps were sampled before writing; they are not persistence timestamps. Reference-access separation relies partly on the process record. Output-schema completeness remains distinct from matching labels and valid witnesses.

The restart's local result note, semantic review, process review, and claims review support this revision but have not been added to R1, R2, A1, or L1. Those links therefore cannot be offered as public reproduction of this new result. No private chats, machine paths, raw runtime logs, or additional internal code are disclosed by this report. Further evidence release requires a separate reviewed package and authorization.

J. 24 OCTOBER LANGUAGE TRANSFER: PROVENANCE AND LIMITS

J.1 Chronology and separation

The fixed answer contract and deterministic core are preexisting work dated 22 October. New L01-L06 prose cases, separate reference answers, and assessment criteria were prepared on 23 October. Execution revision R2 was prepared and neutrally checked on 23 October; the one-shot language-transfer run took place on 24 October. Here execution revision R2 is distinct from the public reading-evidence package named R2. The 23 October V01-V06 restart and its prior abort remain separate historical results. Their original results and evidence remain unchanged and are distinguished in Appendices H-I.

Authorization was bound at 08:04:23.622 UTC on 24 October, with deadline 10:04:23.622 UTC. Six fresh AI roles received neutral instructions before their own case release, without inherited conversation history. All six first captures were sealed at 08:12:09.146 UTC. Four admitted inputs were executed once; processing began at 08:12:20.899 UTC and the durable END was recorded at 08:12:23.873 UTC. The controller observed process exit 0. Separate AI content and technical reviews closed at 08:22:50.6725414 UTC. No research-model/backend call, automatic retry, or mapping repair occurred. AI assistance is not an AI-free study or independent human review.

J.2 What the internal review established

Original prose, actual mapping targets, communication typings, scope reasons, and preserved first-capture bytes were compared with separate references and the fixed contract. Review supported six scope classes and reasons, 16 evaluated labels, eight visible communication typings, and five pairs of opposite-valued compatible witnesses. Two stopped cases have no core labels. No material mapping error or reference dispute was found in this limited review.

Technical review found 28/28 package bindings unchanged, six unique role/release/capture chains, and staging files byte-identical to preserved first captures. Four executable inputs were independently reconstructed from mappings without repeating the core. Four single attempts completed with worker termination confirmed and no finalization failure. This establishes recorded consistency, not complete proof of blindness or semantic correctness. A later documentation-only timestamp-parsing error was corrected before closure; it did not affect the run or its guards and remains recorded in the local closure note.

J.3 Interpretation and evidence access

L01 and L06 allow hidden earlier grants; L03 and L06 allow unobserved acceptance. Their opposing witness worlds preserve visible records and are not additional observations. L02's missing initial grant status is distinct from V06's unresolved authority. L04's grant revocation is distinct from V05's withdrawal of plan acceptance. The new cases and judgments must not replace or inflate the old results.

The six selected synthetic cases remain within two familiar contexts and one stipulated contract. Manual AI mapping and AI review are not independent external validation, automatic natural-language recognition, or a causal improvement study. The known raw answer-schema gap remains: added reviewer explanations must not be credited as core-generated reasons. No result demonstrates hidden coordination, collusive intent, population accuracy, or net safety benefit.

L1 supplies all six synthetic cases, substantive mapping projections, separate references, four complete core result objects, assessment, selected instructions, and internal review excerpts. Original capture/operational records and code remain withheld. Process claims are internally reported, not independently reproduced by L1. It supports Section 6 and Appendix J, not the earlier V01-V06 restart or a general collusion claim. The dated release binding and download verification are recorded in Appendix K.

K. L1 RELEASE AND HISTORICAL REPORT BINDING

L1 is a sanitized reading package, not a runnable reproduction. Viewer access and a download without Google sign-in were checked; internal counterchecking confirmed byte identity. ZIP: QENDROVA-Sprachtransfer-Belege-20261024-L1.zip; 45,912 bytes; 13 files.

ZIP SHA-256: a6aa06c4397c74303443c374faa2720b2b9f5763f540f1c3a92376657aa8e55f

REPORT-MAP.json identifies the preserved QENDROVA-Apart-Report-20261024-Language-Integrated.pdf, not this final report. Its SHA-256 is:

da85eaba10691dcc977f2a6d8925c444abbc39a122a4be4c1aac0a6fbb506bc1

The mapped findings and limits are unchanged here. L1's local-only/not-uploaded statements record preparation before the later authorized release; its bytes remain unchanged. Original hashes do not prove withheld contents. R1/R2/A1 and historical reports remain intact. The L1 access update occurred on 24 October before the original report submission that day. This final report retains those findings and release bindings. Its retrospective synthesis is an interpretation of existing cases, not a new experiment; no new research run or model call was performed.