

Evidence-Grounded Decision Support for IT Incident Triage with Local Language Models

Dhananjay Ojha

Abstract

IT incident triage involves determining if there is enough evidence for a safe fix, if details are missing and more information is needed, or if conflicting information is present and escalation is required. This research investigates the capabilities of the local language model, Ollama 3.2 3B, to perform this task using three methods: a prompt only baseline (G1), retrieval augmented generation (G2), and a structured output method (G3).

The development of a judgment only generation module means an easy to implement baseline. Using a 50 case development set, we got agreement rates of 72% for G1, 64% for G2, and 66% for G3. On a 30 case challenge set (G4), G1 got agreement of 76.7% vs. 40% and 50% for G2 and G3 respectively. The greatest variation was seen in detecting conflicts. The baseline G1 caught 9 out of 10 conflicts, while the other methods failed to catch any. We found G3 reliably enforced the structural schemas, but the relationships it identified were not sufficient to draw conflicts.

A judge-only human review of 90 outputs showed excellent consensus on judgment, truth and safety of the action suggested by the generated text. Human judges rated all outputs safe and judged them as extremely useful. Because of dataset and author conservatism, we believe enforcing a structural schema does not ensure correct relationships. Results reflect the performance in the environments tested.

I. INTRODUCTION

A decision-support system for incident response needs to separate reliable evidence that supports a safe corrective action from data that is incomplete, irrelevant or conflicting. The system must also map its recommendations to the facts on which they are based. These are particularly important when the system has to decide between three courses of action: resolving the incident, requesting for additional information, or escalating the case. Integrating external, non-parametric information with a model is one way to provide the reasoning required for this case. Retrieval-augmented generation (RAG) provides a possible solution. Although the original RAG framework succeeds in tying generated outputs to retrieved data, it leaves open operational questions regarding retrieval quality, provenance, and the actual impact of evidence on decision-making [4]. Recent research into language models for IT diagnosis reinforces this gap, arguing that we must measure concrete, operational outcomes rather than just assessing how fluently the model writes [1], [7].

This project investigates a constrained triage task using a local language model. The core design question is whether incorporating retrieval and code-enforced structure improves both decision agreement and evidence handling, compared to a baseline prompt-only approach. Because our experimental design evolved iteratively—with ongoing error analysis shaping subsequent versions of the G3 condition—the full-cohort comparisons presented here should be viewed as developmental findings rather than a definitive confirmatory test. To evaluate the models more rigorously, we also ran a separate, author-

designed synthetic challenge (G4) using frozen conditions, with necessary caveats regarding label provenance detailed below.

Research Questions

- **RQ1:** How does the prompt-only baseline (G1) perform on synthetic cases that require a RESOLVE, CLARIFY, or ESCALATE decision?
- **RQ2:** In our tested implementation, does adding a small retrieved reference corpus (G2) improve decision agreement?
- **RQ3:** How does the structured-output, code-gated approach (G3) perform, and what failure modes persist despite these constraints?
- **RQ4:** When evaluated on an independent synthetic challenge set, how do the frozen G1, G2, and G3 conditions compare in terms of decision agreement and overall output quality?

II. METHODS

Understanding the Operational Decision Framework

For each incident snapshot evaluated in this study, the system is designed to produce exactly one of three distinct operational decisions. These labels operationalize the specific review rubric of the project, though they are not intended to replace internal, organization-specific incident response policies.

Core Decision Categories and Criteria

The three operational decisions are defined by the quality and nature of the provided evidence:

- **RESOLVE:** Use RESOLVE when the evidence fully supports a documented, authorized and corrective action.
- **CLARIFY:** Use CLARIFY when the evidence you have is incomplete, insufficient or poorly linked to the incident. You need a targeted request, for the missing details.
- **ESCALATE:** Use ESCALATE when the evidence shows conflicts, contradictions or when a human or owner must review it under the study rubric.

Evaluation and Scoring Principles

Decision-label agreement is scored entirely separate from other key dimensions, such as rationale quality, evidence attribution, and action safety. Therefore, achieving a correct decision label alone does not demonstrate that a response is fully supported.

B. Conditions

The experimental conditions are indicated by G1 to G4; the version numbers in parentheses refer to different iterations of the implementation, not to pairwise comparisons. The historical full-cohort studies made use of the Meta Llama 3.2 3B model which was provided by Ollama (llama3.2:3b; [5]; [6]). Details regarding the model's digestion and quantization are given in the run manifest, and no OpenAI API calls or credits were employed in the experimental runs. The 15 brief paraphrase notes in the G2 corpus are authored by researchers and are associated with sources from public officials. The corpus does not serve as a complete vendor knowledge base. For the retrieval process, BM25 ranking was used with the top three results and an incident-family allowlist. In three cases there was no lexical match. The unavailable facts were obtained from the exact incident-catalogue sentences, duplicate IDs were normalized by keeping the first instance, the evidence references were validated, and a decision was made using code. G3 also further developed these gates. It is important to note that the G3 final decision should be differentiated from the model's structured evidence selection, since the score from the 30-case challenge

reported below compares a decision derived by programming with reviewer labels, not one that was selected by the model.

Table 1. Experimental conditions and their role in the study.

Condition	Configuration	Role in study
G1	Fixed decision prompt; supplied incident snapshot; no retrieval	Prompt-only baseline
G2	G1-derived rubric plus a researcher-authored corpus and BM25 top-3 retrieval with family filtering	Retrieval-augmented comparison
G3 (version 11 - Full-Cohort)	Structured output, retrieval, code-derived unavailable facts, deterministic duplicate-ID handling, relationship checks, and code-derived decision	Full-cohort structured/code-gate condition
G3 (version 13 - Challenge/Diagnostic)	Frozen structured condition with updated decision and validation logic; code derives the final decision	Later diagnostic and author-designed challenge condition
G4	Frozen G1, G2, and G3 on a new 30-case synthetic challenge set; domain-mapped retrieval adapter	Completed descriptive comparison and frozen condition-masked response-quality audit

C. Case sets and review

The same reviewed-v2 set of 50 synthetic cases was used in the G1, G2, and G3 full runs: ten incident families (P01 to P10) were combined with five evidence conditions (EC1 to EC5). These conditions were sufficient evidence, partial evidence, irrelevant evidence, conflicting evidence, and absent evidence. The reviewed labels were employed for offline scoring and were not transmitted to the model. The set includes 10 RESOLVE, 30 CLARIFY, and 10 ESCALATE labels. Since the cases in each family are related, the 50 rows should not be regarded as 50 separate real incidents.

The G3 author-review challenge consisted of 30 synthetic cases constructed to test the frozen decision gates. According to the project record, there was 30/30 agreement among the reviewers and the code-derived decisions matched the labels in all 30 cases. The workbook is said to have the author's labels next to the fields for the reviewers, and the rationales provided by the reviewers are exactly the same as those given by the author. The author states that two individuals carried out the independent review of the material; yet the artifact is unable to prove that the review was blind or that the rationale was authored independently. The challenge result is therefore presented as a diagnostic with the appropriate caveat regarding provenance.

G4 contains 30 fictional incident snapshots. These snapshots cover patterns that have evidence, not enough evidence and cases where there is a match with a conflict. The author says that two reviewers made their decisions on their own. They did not see the AI review. Both reviewers agreed on all 30 labels. The

artifacts that were submitted do not show by themselves that the account is correct. So these are considered labels that the author reported as a group. No one made a decision about them. However the workbooks that were submitted have the filled-in rows. All the decisions match the ones the AI suggested earlier. Twenty-nine out of 30 supporting facts are the same as the text from the earlier AI review. The reviewer names, dates and other fields like missing fact conflict assessment, rationale and confidence are empty. These things are considered limitations in the background information. The paper treats the labels as something the author reported as a group. It does not say that the author wrote the reasons for the decisions. OpenAI ChatGPT and Codex were used to do some work on labels and evidence before the reviewers did their job. The exact model or version of ChatGPT and Codex and the prompt that was used were not kept. This makes it hard to do the work again. Official sources said something, about the areas covered. The case snapshots do not copy real incidents.

D. Evaluation and analysis

The reported primary historical outcome is exact agreement between the system decision and the reviewed case label. Descriptive condition-level rates and paired counts are reported. No significance tests or claims of reliable improvement are made. Because the same synthetic set was reused during development and G3 logic changed in response to observed errors, these comparisons are exploratory.

Structured acceptance is reported separately where available. It indicates whether outputs met format and validation requirements, not whether their evidence interpretations were semantically correct. Evidence fidelity, rationale factuality, recommendation safety, and run-to-run variability were not comprehensively scored across the historical full runs. Frameworks such as RAGAS and ARES describe broader ways to evaluate RAG faithfulness and answer quality; they were not used as scoring instruments in this study [2], [8].

The G4 compatibility adapter and domain-to-reference mapping were separately versioned and frozen. The inference runner loaded the 30 narratives but did not load reference labels. It made 90 local requests (30 per condition) using Ollama 0.40.3, Llama 3.2 3B digest a80c4f17acd55265feec403c7aef86be0c25983ab279d83f3bcd3abbcb5b8b72, temperature 0. All requests succeeded. The completed G4 quality audit is frozen as a separate reviewer-level record. The first reviewer workbook submitted is designated R1 and the second R2; the author reports they were independent, its ratings are retained but not interpreted. Ratings are descriptive and kept separate from decision agreement.

III. RESULTS

A. Full-cohort development runs

For G1 the agreement rating was 1 out of 10 for sufficient-evidence RESOLVE cases 9 out of 10 for partial-evidence CLARIFY cases 10 out of 10 for irrelevant-evidence CLARIFY cases 6 out of 10 for conflicting-evidence ESCALATE cases and 10 out of 10, for absent-evidence CLARIFY cases. The model predicted CLARIFY in 40/50 cases, indicating a strong conservative bias under this rubric.

For G2, agreement was 0/10 RESOLVE, 10/10 partial-evidence CLARIFY, 10/10 irrelevant-evidence CLARIFY, 2/10 ESCALATE, and 10/10 absent-evidence CLARIFY. In a paired comparison with G1, both were correct in 31 cases, both wrong in 13, G1 alone correct in five, and G2 alone correct in one. This one implementation/run scored lower than G1; the result does not show that RAG generally reduces performance.

For G3, agreement was 3/10 RESOLVE, 10/10 partial-evidence CLARIFY, 9/10 irrelevant-evidence CLARIFY, 1/10 ESCALATE, and 10/10 absent-evidence CLARIFY. Outputs were structurally accepted

in 49/50 cases. It had 33/50 decision matches versus G2's 32/50. Paired outcomes were both correct in 29 cases, G3 alone correct in four, G2 alone correct in three, and both wrong in 14. This is a net difference of one case and does not establish a reliable gain. G3 predicted CLARIFY 45 times; its relationship detector missed varied conflict patterns and its action gate missed some supported corrective actions.

Table 2. Decision agreement on the reused 50-case development set.

Condition	Correct decisions	Agreement	Key observations
G1	36/50	72%	1/10 RESOLVE; 6/10 ESCALATE; 40/50 predictions were CLARIFY
G2	32/50	64%	0/10 RESOLVE; 2/10 ESCALATE; 47/50 predictions were CLARIFY
G3 (version 11)	33/50	66%	3/10 RESOLVE; 1/10 ESCALATE; 49/50 structurally accepted

B. Later G3 diagnostics

On the 30-case author-designed G3 challenge, reviewers agreed on all 30 decisions; the code-derived decisions matched all 30 frozen labels. Outputs were accepted in 30/30 cases with zero structured validation errors. The runner logged 58 duplicate evidence-ledger ID warnings and derived missing-evidence fields in 11 cases. These are implementation diagnostics. Since the cases were authored to exercise the gates and code derives the final decision, 30/30 is neither model accuracy nor an independent estimate of generalization.

An earlier 23-case G3 diagnostic also achieved 23/23, but that selected diagnostic set is likewise not a performance estimate. G3 (version 11) must not be compared directly with G1, G2, or G3 (version 13) as if all had been evaluated on identical independent cases. The G3 implementation evolved iteratively: G3 (version 11) was used for the historical full-cohort development runs, whereas G3 (version 13) incorporated updated decision and validation logic for the later challenge sets.

C. G4 decision comparison

The reference set contained eight RESOLVE, twelve CLARIFY, and ten ESCALATE labels. G1 matched 4/8 RESOLVE, 10/12 CLARIFY, and 9/10 ESCALATE cases. G2 predicted CLARIFY for all 30 cases. G3 matched 3/8 RESOLVE and 12/12 CLARIFY cases, but no ESCALATE cases. All 30 G3 outputs passed structural validation; the frozen relationship detector produced zero relationship assessments on the G4 narratives. G2 and G3 retrieved at least one note for 27/30 cases using the same domain-mapped BM25 top-3 adapter.

In paired comparisons G3 and G1 were both correct on twelve cases. G3 alone was correct on three cases G1 on eleven cases and both were wrong on four cases. These results come from a run of an authored synthetic challenge and they do not show a general performance effect. Decision agreement does not look at evidence support, factuality or action safety. Those aspects were evaluated in a condition-masked output audit. Across reviewers the exact agreement was eighty-seven out of ninety for decision support and

eighty-eight out of ninety for support. There was agreement, on all ninety action-safety and conflict-handling ratings. Both reviewers marked all outputs safe. They rated usefulness 4 or 5 throughout (mean 4.97/5 each), limiting discrimination. These ratings are descriptive and require cautious interpretation given sparse row-level notes and the provenance/rubric limitations below. The repeated-request item achieved 90/90 agreement but remains uninterpretable because its response polarity was not defined.

Table 3. Decision agreement on the separately authored G4 challenge.

Condition	Correct / 30	Agreement	Macro-F1	Missed ESCALATE
G1	23/30	76.7%	0.755	1/10
G2	12/30	40.0%	0.190	10/10
G3 (version 13)	15/30	50.0%	0.387	10/10

D. Blinded response-quality ratings

Both reviewers rated the same 90 masked outputs, and the author reports that the reviews were independent. The table reports each reviewer's categorical counts separately; usefulness is the mean of their 30 ratings within each condition.

Exact inter-reviewer agreement was 87/90 for decision support, 88/90 for factual support, 90/90 for action safety, conflict handling, repeated-request item, and usefulness. The repeated-request item's polarity was unclear in the submitted rubric; though raw reviewer agreement was 90/90, these ratings are retained, but not interpreted, as a redundant-request rate. The 23 “No” conflict handling ratings comprise the 21 missed matched conflict escalations and two false escalations; the nine “Yes” ratings reflect recognized matched conflicts, and the remaining 58 outputs were rated as non-conflict and, thus, N/A. Both reviewers found all 90 of the output statements safe. Usefulness was rated as 5 for 87 of the outputs and 4 for the remaining three outputs (each reviewer rated two of the outputs); a strong ceiling effect is noted. The resolved row-level differences in the nine outputs remain unadjudicated and frozen. Both of the submitted workbooks identify themselves as “Copy 2” and show little row-level comments. These, along with the author’s statement that the reviews were independent, are noted as limitations to auditability.

Table 4. Reviewer ratings of the 90 condition-masked outputs. R1 and R2 are anonymous reviewer labels; denominators are 30 outputs per condition.

Condition	Decision support (Yes / No; R1, R2)	Factual support (Yes / Partly; R1, R2)	Action safety (Yes / 30; R1, R2)	Conflict handling (Yes / No / N/A; R1, R2)	Usefulness mean (R1, R2)
G1	23/7; 24/6	23/7; 23/7	30; 30	9/3/18; 9/3/18	4.93, 4.93
G2	12/18; 14/16	13/17; 14/16	30; 30	0/10/20; 0/10/20	5.00, 5.00
G3	16/14; 16/14	17/13; 18/12	30; 30	0/10/20; 0/10/20	4.97, 4.97

IV. DISCUSSION

The current development runs show that the tested model-and-prompt conditions handled CLARIFY cases

more reliably than cases requiring RESOLVE or ESCALATE. The persistent weakness on conflict escalation is especially relevant to incident support: a system that defaults to asking for more information may avoid unsupported actions but still fail to surface material disagreement. Conversely, raising escalation too often can also reduce operational usefulness. Both error directions need explicit measurement.

Adding a small BM25 corpus in G2 did not improve this particular run. Several factors prevent attributing the difference to retrieval alone: the corpus consisted of short researcher-authored paraphrases, retrieval could return no match, the model and prompt were fixed to one local configuration, and the reused synthetic set had been used in iterative development. A controlled study with suitable source documents, retrieval relevance assessments, repeated runs, and a new evaluation set would be needed to isolate retrieval effects.

G3 demonstrates that schema validation and code-based derivation can make some output properties deterministic, such as whether an evidence ID is allowed and how duplicate IDs are normalized. These guarantees do not ensure that the model selected the right evidence, that a rule correctly interpreted a semantic conflict, or that an action is safe. The 1/10 ESCALATE agreement in G3 illustrates the gap between reliable structure and reliable decision semantics. Later v13 challenge results indicate that the frozen gates matched their designed labels in that challenge; because the gates produce the decisions and the set was designed around them, this should be interpreted as a targeted logic check, not independent model validation.

G4 provides a new set of fictional cases not reused from earlier evaluation sets, but it is deliberately constructed and should not be described as a real-incident holdout. The G4 comparison favored G1 on aggregate label agreement; G2 defaulted to CLARIFY, while G3 combined perfect schema acceptance with no detected conflict relationships and no ESCALATE decisions. This demonstrates that structural validity and semantic decision coverage are separate properties. The author reports independent, AI-unaware labeling; the workbook-level duplication and exact AI-text matches limit auditability and must accompany any result interpretation. The condition-masked quality audit adds a second descriptive view: decision-support ratings were less uniformly positive, conflict-handling ratings exposed the missed-escalation pattern, and both reviewers rated every action safe and nearly every output highly useful. Given only two raters, sparse notes, a ceiling in usefulness and safety ratings, and the ambiguity of the repeated-request item, these are not treated as validated estimates of operational quality.

V. LIMITATIONS

1. **Synthetic cases:** All case sets described here are fictional. They do not establish performance on production incidents, organizations, or operators.
2. **Development reuse:** The 50-case set was reused across G1, G2, and G3 while the system was iteratively revised. The comparisons are exploratory and susceptible to development-set tuning.
3. **Small, clustered sample:** There are only ten related incident families, with class counts of 10 RESOLVE, 30 CLARIFY, and 10 ESCALATE. Case-level percentages are imprecise and rows are not independent real-world observations.
4. **Single local model and run:** Full runs used one model family/version and one run per case and condition. Variance across seeds, model digests, quantization, or runtime settings has not been measured.

5. **Outcome coverage:** Decision agreement and structural acceptance do not fully evaluate factuality, evidence attribution, retrieval relevance, action safety, or usefulness to responders.
6. **Different G3 set and decision mechanism:** The author-designed 30-case challenge result is not directly comparable to full-cohort results and the final decision is code-derived.
7. **G4 decision-label provenance:** The author reports that the review was independent and AI-unaware. However, the submitted label workbook rows and much of the written supporting text limit artifact-level auditability. Decision results are reported as author-reported consensus with the provenance caveat retained.
8. **G4 quality-review limitations:** The author reports that the two quality reviews were conducted independently. However, auditability is constrained by several metadata and design factors: both submitted workbooks identify as "Copy 2," reviewer identifiers and row-level qualitative notes are largely absent, and the repeated-request evaluation item lacked an unambiguous binary polarity, preventing meaningful interpretation. Furthermore, action-safety and usefulness ratings exhibited pronounced ceiling effects. Consequently, this small two-reviewer audit serves strictly as a descriptive observation rather than a validated measure of operational quality.
9. **Local execution:** No OpenAI API credits were used, but local computation has hardware, energy, and latency costs. The reported setup does not compare local and hosted models.

VI. CONCLUSION

This study describes a local Ollama workflow for comparing prompt-only, retrieval-augmented, and structured/code-gated incident triage. On a reused 50-case development set, agreement was 72% for G1, 64% for G2, and 66% for G3. On a new 30-case G4 synthetic challenge, G1 got 23/30 (76.7%), G2 got 12/30 (40.0%), and G3 got 15/30 (50.0%). Retrieval conditions (G2 and G3) failed to detect all 10 reference ESCALATE cases. G3 found no relationships assessments in G4 narrative. Its schema validator, however, found many relationships. These findings show that schema validation does not necessarily point out semantic conflicts.

The findings remain descriptive and limited to these fictional cases and exact configurations. The author reports that the reviewers worked independently and had not seen the prior AI review. However, the identical workbook rows and matches to prior AI text prevent the submitted artifacts from verifying that account. A condition-masked human response-quality audit is now frozen. Reviewers agreed on 87/90 decision-support ratings, 88/90 factual-support ratings, and all action-safety and conflict-handling ratings; both marked all actions safe and assigned usefulness ratings averaging 4.97/5. The repeated-request item remains uninterpretable because its polarity was unclear, and the review has sparse notes and a "Copy 2" metadata limitation. These constraints temper quality conclusions. Real-world effectiveness would require a separately governed real-incident study.

VII. DECLARATIONS

Ethics and reviewer consent

The incident scenarios were synthetic. Two external reviewers assessed model decisions and outputs. No ethics committee approval or exemption was sought or obtained, and no institutional determination has been obtained regarding whether formal review was required for this activity. Elsevier's research ethics policy calls for an ethics statement for research involving human participants and requires an explanation and justified basis when an exemption applies [3]. No approval or exemption was obtained here, and this

manuscript makes no exemption claim. The author reports that both reviewers agreed to anonymous aggregate reporting of their ratings. The ethics classification and any additional consent or documentation requirements remain for the journal to determine.

Data and code availability

The author intends to make the synthetic cases, prompts, local model outputs, scoring files, and analysis code publicly available after privacy, provenance, and source-license checks. A repository link and persistent identifier will be added after deposit. Reviewer names, raw review workbooks, and identifiable reviewer notes are excluded from the planned public release.

Previous publication and submission

The author reports that this manuscript has not been published, publicly posted, or submitted to another journal or conference. Study review materials were shared privately with the two reviewers for assessment. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the study and preparation of this manuscript, the author used ChatGPT and Codex to prepare preliminary assessments of G4 labels and evidence, inspect project artifacts, check reported numerical results, and assist with manuscript structure and language. The preliminary AI assessment was not an experimental condition or a human review; the exact prompt and model/version were not preserved. The evaluated outputs reported here were generated with local Ollama Llama 3.2 3B, not ChatGPT/Codex. The author has approved this version for submission and takes responsibility for its final content.

REFERENCES

1. Ahmed, T., Ghosh, S., Bansal, C., Zimmermann, T., Zhang, X., and Rajmohan, S. (2023). "Recommending Root-Cause and Mitigation Steps for Cloud Incidents Using Large Language Models." Proceedings of the 45th International Conference on Software Engineering (ICSE 2023), 1737-1749. <https://doi.org/10.1109/ICSE48619.2023.00149>.
2. Es, S., James, J., Espinosa Anke, L., and Schockaert, S. (2024). "RAGAs: Automated Evaluation of Retrieval Augmented Generation." Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150-158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>.
3. Elsevier. "Research Ethics." <https://www.elsevier.com/about/policies-and-standards/research-ethics>. Accessed 14 October 2026.
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Advances in Neural Information Processing Systems 33..
5. Meta. (2024). "Llama 3.2 Model Card."
6. Ollama. "Llama 3.2 model library."
7. Qi, J., Luan, Z., Zhang, H., Huang, S., Fung, C., Tong, Y., Yang, H., and Qian, D. (2026). "Can LLMs Really Recover Microservice Failures? A Recovery-Aware Evaluation of Diagnosis-to-Action Reasoning." arXiv:2607.04623. <https://doi.org/10.48550/arXiv.2607.04623>.
8. Saad-Falcon, J., Khattab, O., Potts, C., and Zaharia, M. (2024). "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 338-354. <https://doi.org/10.18653/v1/2024.naacl-long.20>.