Proceedings of 5th International Conference CNDGS'2026
ISSN 2669-2023 (print) / ISSN 2538-8959 (on-line) / **https://doi.org/10.47459/cndcgs.2026.16**

Engineering Governable Agentic Workflows with Large Language Models for Defence Decision Support: A Conceptual Framework

Antonín ROUSEK[1]

[1]*Institute of Intelligence Studies, University of Defence, Kounicova 156/65, 662 10 Brno, Czech Republic*

Correspondence: [1]antonin.rousek@unob.cz

Abstract

This paper examines how large language models can be embedded in governable *agentic workflows* for defence decision support. It argues that reliability and accountability in such systems depend on workflow architecture, not only on model capability. Building on the author's earlier I→E→R model, the paper proposes an audit-first governance framework with three layers: explicit task decomposition, evidence grounding with provenance, and bounded control with logging and authorization gates. The framework is specified through five inspectable roles, four design variables, and *bounded autonomy* as the guiding design principle. The contribution is conceptual: it defines conditions under which agentic analytical processes can remain transparent, traceable, reconstructable, and auditable, and prepares the framework for later case-study or proof-of-concept evaluation.

KEY WORDS: *agentic workflows; auditability; defence decision support; large language models; workflow engineering*

Citation: ROUSEK, A. Engineering Governable Agentic Workflows with Large Language Models for Defence Decision Support: A Conceptual Framework. In Proceedings of the Challenges to National Defence in Contemporary Geopolitical Situation, Brno, Czech Republic, 7-10 September 2026. ISSN 2538-8959, https://doi.org/10.47459/cndcgs.2026.16

1. Introduction

Large Language Models (LLMs) are increasingly used as the reasoning engine of *agentic workflows*: multi-step pipelines in which a model coordinates planning, retrieval, tool use, and iterative refinement with partial autonomy over sequencing [1, p. 3; 2, p. 2, 18]. In defence decision support, this creates a procedural problem. Analytical outputs used under uncertainty and command responsibility must remain inspectable, attributable, and open to intervention, yet agentic workflows distribute error across planning, evidence selection, tool calls, synthesis, and review.

Model-level risks are already familiar: LLMs can generate unsupported or overconfident claims [3, p. 5; 4, p. 7]. Agentic workflows add workflow-level risks, including cumulative uncertainty, tool-mediated failures, and broader attack surfaces [5, p. 2; 6, p. 3]. Human oversight is not a complete answer unless the process exposes errors in a form that operators can detect and contest [7, p. 8–9, p. 3]. For defence organizations, this makes workflow design a governance issue, because review, responsibility, and timely intervention depend on records of how an output was produced [10, p. 3; 11, p. 39].

This paper addresses that governance gap. It builds on the author's I→E→R model [12], which linked Integration, externalization, and Process Rigour in agentic analytical work, but did not specify a defence-oriented governance architecture. The contribution here is an audit-first framework that translates externalization into three layers, five inspectable roles, four design variables, and a bounded-autonomy principle.

The risk landscape motivating the framework is summarized in Figure 1. Section 2 positions the argument in the relevant literature. Section 3 explains the constructive synthesis method. Section 4 presents the framework and an illustrative workflow. Section 5 evaluates the framework, draws implications, and states limitations.

Peer-review under responsibility of General Jonas Žemaitis Miltitary Academy of Lithuania and University of Defence, Czech Republic

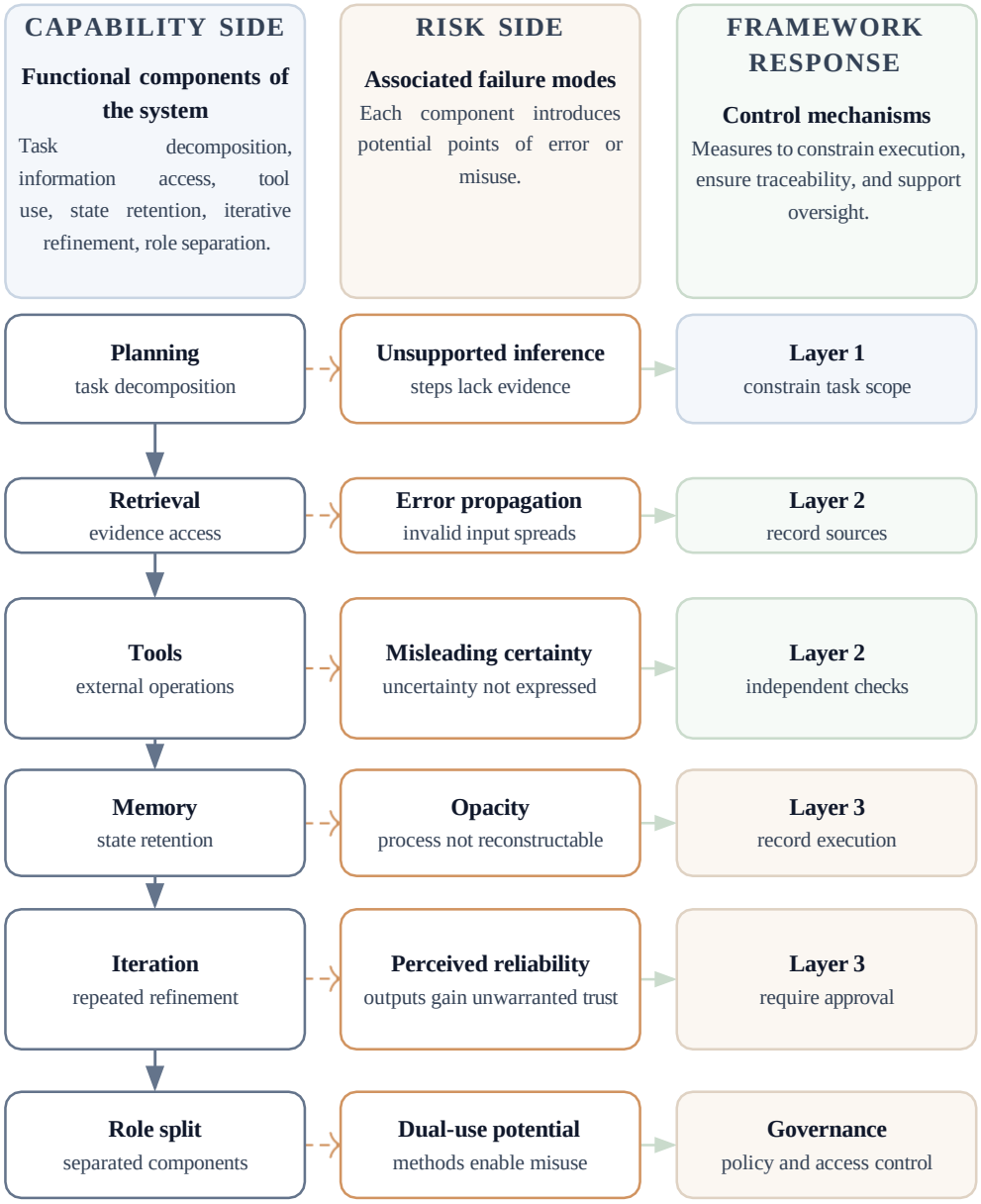

Fig. 1. **Risk landscape for agentic workflows**. The figure maps workflow capabilities to failure modes and to the control layers proposed in this paper. Source: Author.

2. Background

The analytical starting point is the I→E→R model [12]. In that model, Integration denotes the embedding of agentic components into analytical procedure, externalization the capture of intermediate reasoning artefacts, and Process Rigour the discipline of the final product. The model explains why externalization matters; this paper specifies how it can be operationalized for defence workflows.

Three literature streams inform the framework. Work on agentic architecture supplies the vocabulary of planning, retrieval, tool use, memory, and coordination [1; 2; 13, p. 2]. Retrieval-grounded reasoning and provenance literature motivate explicit evidence binding and caution that provenance alone does not guarantee evidential adequacy [14, p. 2–16, p. 2]. Governance and assurance literature defines the institutional constraints of risk management, oversight, traceability, and lifecycle control [10, p. 3; 11, p. 39; 17, p. 5; 18, p. 4].

3. Methods

This paper uses constructive conceptual synthesis. The procedure is not a systematic literature review; it draws requirements and component vocabularies from existing work and integrates them into a governance specification. The resulting claim is conceptual: the framework identi-fies conditions for governable defence-oriented workflows but does not show empirically that compliant systems produce better outputs.

The synthesis proceeds in four steps. First, it defines the defence decision-support problem in terms of workflow-level failure modes and governance requirements. Second, it derives design requirements from the three literature streams identified above. Third, it translates those requirements into the audit-first framework, mapping each requirement to a layer, role, or design variable. Fourth, it evaluates the result against three criteria.

Coherence asks whether the layers, roles, and variables form a consistent whole. Completeness asks whether the framework covers the derived requirements, with gaps noted where coverage is partial. Plausibility asks whether the control logic fits documented governance structures, especially NATO AI principles [10, p. 3], the EU AI Act requirements for high-risk systems [18, p. 4], and UK defence AI assurance guidance [11, p. 39].

The evaluation does not involve workflow execution, comparative case analysis, or expert elicitation. It is intended to make later empirical testing more precise.

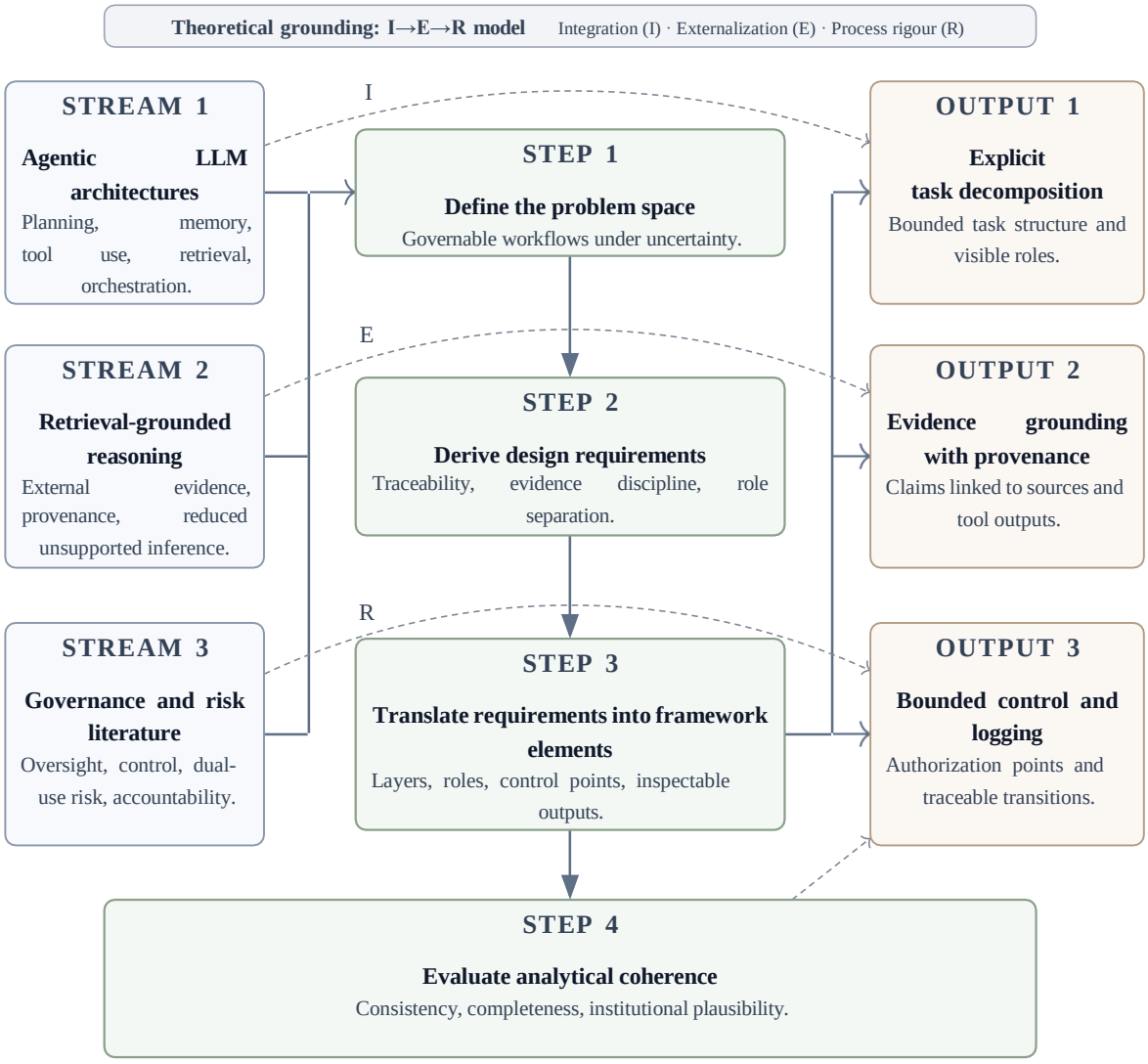

Fig. 2. **Constructive synthesis procedure.** The figure summarizes how the I→E→R model and three literature streams are translated into design requirements, evaluation criteria, and the audit-first framework.

4. Results

This section presents the audit-first framework and an illustrative workflow. Both are conceptual outputs of the synthesis procedure; neither is an empirically validated system.

4.1 Audit-First Framework

The framework translates the externalization emphasis of the I→E→R model [12] into a three-layer governance architecture. Figure 2 presents the layers, roles, design variables, and authorization logic; Figure 3 later shows an end-to-end instantiation.

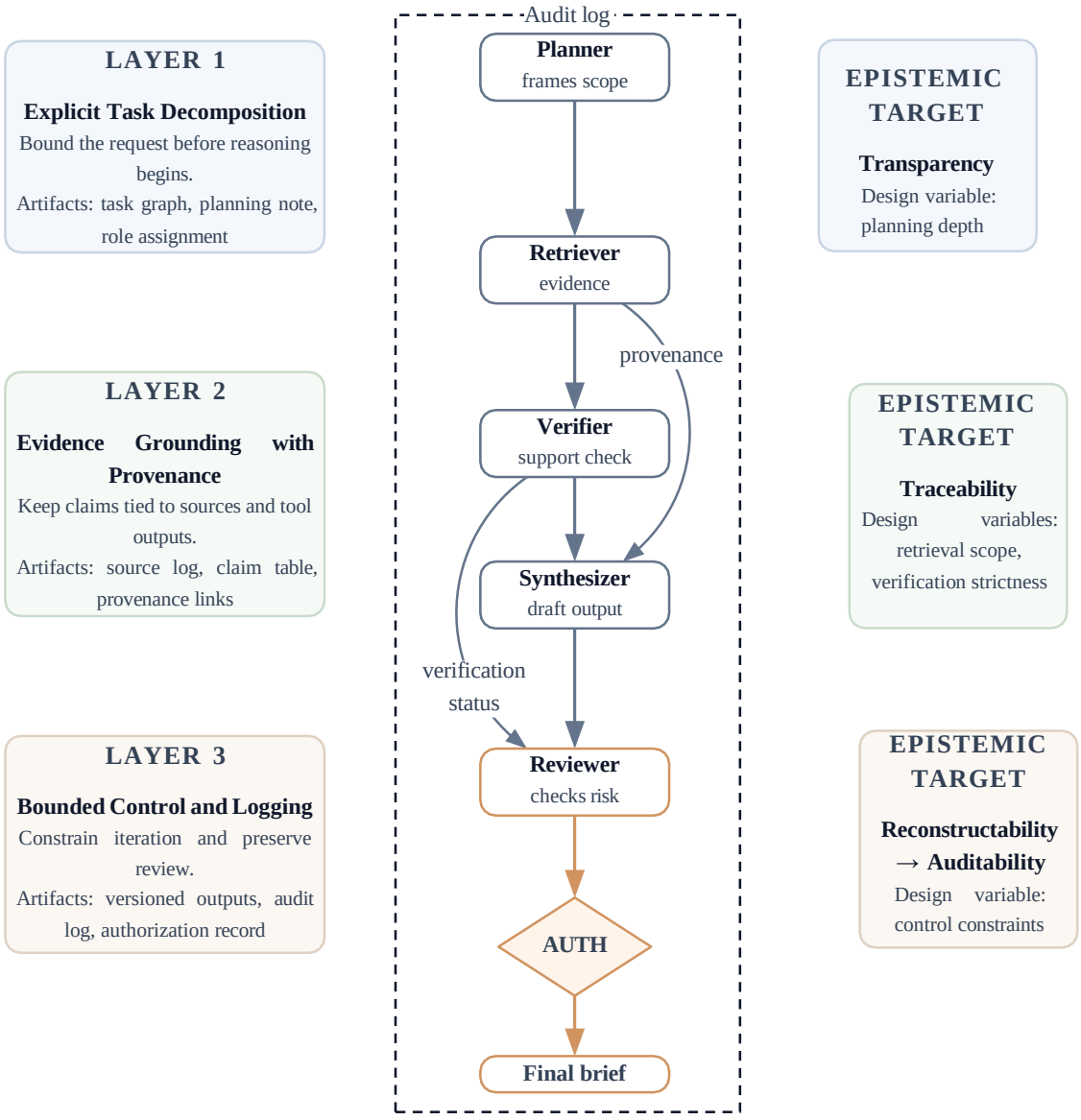

Fig. 3. **Three-layer audit-first framework for governable agentic workflows.** The figure shows the three control layers, five roles, four design variables, and authorization gate.

The framework has three dependent layers. Layer 1, explicit task decomposition, turns a broad request into a task graph of sub-problems, required outputs, and role assignments. Layer 2, evidence grounding with provenance, requires generated claims to be linked to retrieved documents, tool outputs, or other inspectable artefacts. Layer 3, bounded control and logging, records sequencing, escalation, authorization, and release decisions. Layer 1 fixes the scope for Layer 2; Layer 2 produces the provenance records that Layer 3 logs and gates.

Reasoning is distributed across five inspectable roles. The Planner frames the task and produces the task graph. The Retriever gathers evidence for specified sub-questions. The Verifier marks claim as supported, contested, or unsupported. The Synthesizer drafts the analytical product from verified artefacts. The Reviewer checks the draft against the task graph, evidence coverage, and unresolved uncertainty before authorization. This separation makes failures easier to localize than in a monolithic pipeline.

Reasoning is distributed across five inspectable roles. The Planner frames the task and produces the task graph. The Retriever gathers evidence for specified sub-questions. The Verifier marks claim as supported, contested, or unsupported. The Synthesizer drafts the analytical product from verified artefacts. The Reviewer checks the draft against the task graph, evidence coverage, and unresolved uncertainty before authorization. This separation makes failures easier to localize than in a monolithic pipeline.

Four design variables govern strictness. Planning depth controls decomposition granularity. Retrieval scope controls the breadth and type of admissible sources. Verification strictness sets the evidential threshold for accepting or contesting claims. Control constraints define iteration limits, escalation rules, artefact requirements, and authorization checkpoints. Numeric thresholds are deployment-specific; the framework fixes the control logic.

The framework uses four related criteria from [12]. Transparency means that a reader can follow the reasoning in the output. Traceability means that each assertion can be linked to a source or evidence item. Reconstructability means that the inference chain can be assembled later from stored artefacts. Auditability is the institutional form of reconstructability: an authorized reviewer can reconstruct a workflow execution under defined retention, provenance, and access rules [19; 20, p. 8]. Auditability is the binding constraint; the other three properties are necessary conditions.

Bounded autonomy is the principle that agentic automation should be limited to the minimum needed for analytical reach while keeping the process reconstructable and under meaningful human control. It is operationalized through maximum iteration counts, mandatory authorization checkpoints, required artefact generation, provenance coverage thresholds, and escalation thresholds. In Figure 2, the AUTH decision diamond represents the final release gate.

The evidence layer treats provenance as necessary but insufficient. A provenance record links a generated claim to the source or tool output from which it was derived. In line with W3C PROV, such records support later assessment of quality, reliability, and trustworthiness [19, p. 1]. However, retrieval-grounded systems may still miss relevant evidence or rely on outdated material [15, p. 1; 16, p. 2]. Reviewers should therefore treat provenance as a condition of accountability, not as proof that the evidence base is complete or unbiased.

As an illustration, consider a request asking whether political developments indicate a shift in a country's alignment toward a rival bloc. The Planner separates events, official positions, and independent assessments. The Retriever logs government statements, news reports, and policy commentary. The Verifier marks a reported public statement as supported but marks a formal policy shift as contested because no policy document is available. The Synthesizer carries forward the supported claim and uncertainty; release occurs only after review and authorization. The point is not the scenario itself, but the prevention of a plausible unsupported inference from silently entering the product.

1.1 Illustrative Workflow

Figure 4 expands the framework into an end-to-end workflow from task intake to released brief. It is an illustrative design, not a universal reference model or existing system. The lanes separate Human Authority, Agentic Workflow, Tools, and Storage/DB; solid arrows show process flow and dashed arrows show artefact storage¨

The workflow proceeds in six stages: task intake, planning and scope approval, retrieval and provenance capture, verification and escalation, synthesis and review, and release authorization. Each stage produces a named artefact, such as a session log, task graph, source log, claim table, draft artefact, review record, or authorization record. These artefacts persist independently of the execution and can be assembled into an audit trail.

The influence of structured analytical tradecraft is limited to three analogues: explicit task framing, evidence testing before synthesis, and independent review at consequential transition points [21, p. 311; 22, p. 18]. The workflow adapts these principles to an agentic computational setting by making intermediate artefacts and human authorization points persistent.

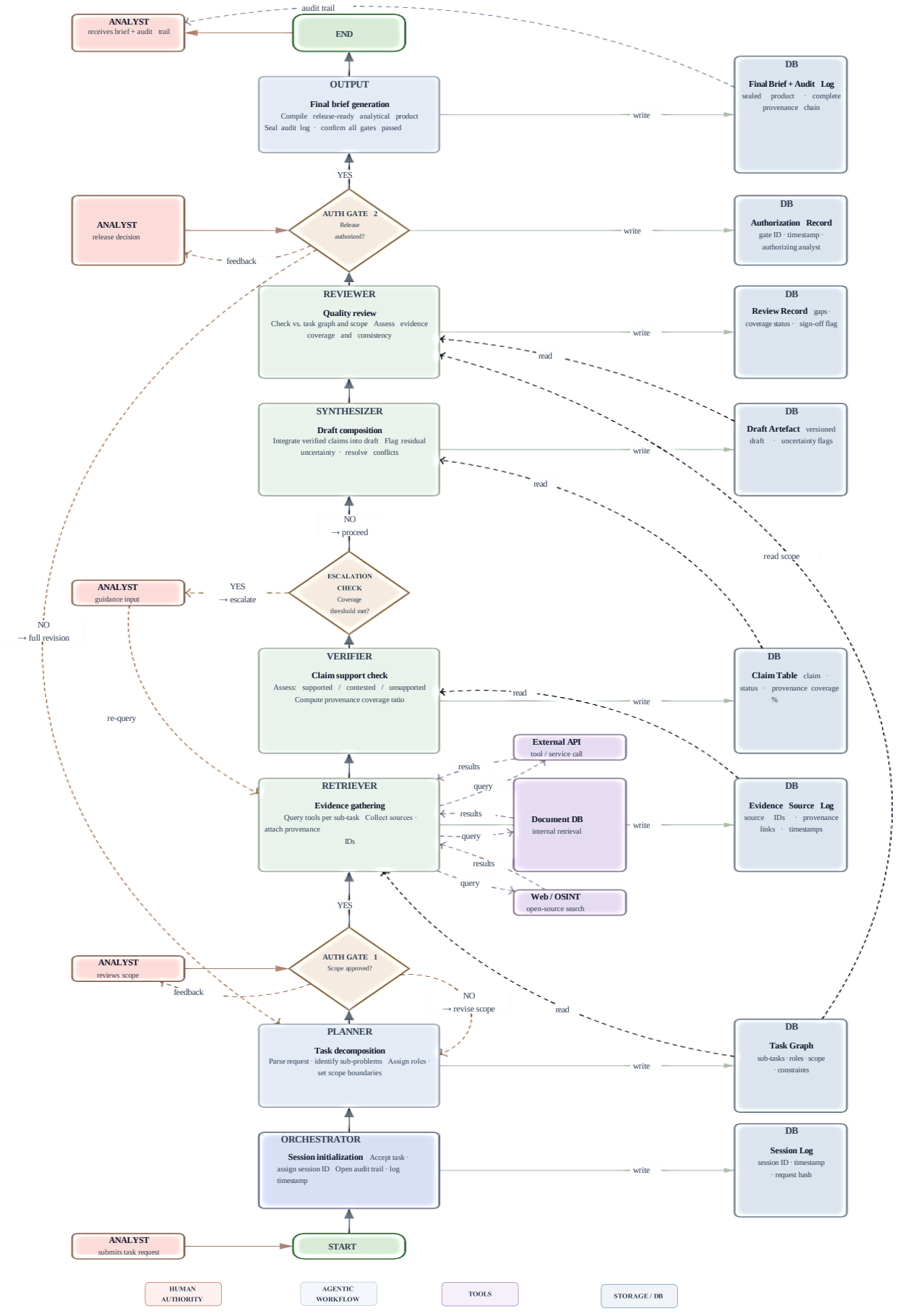

Fig. 4. **Analyst-to-brief workflow.** The figure shows how roles, artefacts, storage, and authorization gates can be arranged in an inspectable analytical process.

5. Discussion

This section evaluates the framework conceptually, draws implications for practice and research, and states limitations.

5.1. Framework Evaluation

The framework is coherent because its layers address distinct governance functions: scope, evidence, and control. The five roles have separable functions and artefact outputs, while the four variables map to the layers without redundancy. The dependency relation is also consistent: evidence handling presupposes a task graph, and controlling logging presupposes provenance records.

It is complete with respect to the derived requirements. Agentic architecture concerns are covered by role decomposition and planning depth. Retrieval and provenance concerns are covered by the evidence layer and verification strictness. Governance concerns are covered by bounded control, logging, and authorization. One partial gap remains: the framework requires provenance coverage thresholds but does not define a method for calculating them. It is plausible against the institutional reference standard. Human authorization and traceability fit NATO AI principles [10, p. 3]. Bounded autonomy and logging align with EU AI Act requirements for high-risk systems [18, p. 4]. Role separation and artefact retention are consistent with UK defence AI assurance guidance on review chains and audit evidence [11, p. 39]. No structural incompatibility was identified, although local deployment would require adaptation.

5.2. Implications for Practice and Research

The first implication is practical: reliability should be assessed as an architectural property. Recent evaluation work suggests that error rates and uncertainty depend on how a workflow is structured, observed, and constrained [6, p. 3; 23, p. 9]. Planning depth reduces scope ambiguity, provenance exposes support relations, and control constraints limit unchecked propagation of earlier mistakes. The assessment question therefore shifts from how capable the model is to how is the workflow designed.

The second implication is organizational: role separation supports fault localization and mean-ingful human control. A logged distinction between planning, retrieval, verification, synthesis, and review makes it possible to identify where a failure entered the pipeline [2; 13, p. 2]. Effective oversight also requires intelligible artefacts, intervention points, and authority to interrupt or redirect the process [9, p. 3; 11, p. 39; 18, p. 4; 24, p. 6]. This creates a governance-throughput trade-off. Manual review maximizes control but limits speed, while maximal automation increases throughput but weakens institutional defensibility. The target state is therefore bounded autonomy: enough orchestration to support analysis, with architecturally enforced gates at scope approval, synthesis, and release. The third implication is methodological: auditability cannot be added after output generation. Audit literature emphasizes process evidence, documentation, and independence over persuasive narrative alone [20, p. 8]. In this framework, auditability arises from the joint presence of task graphs, provenance records, logs, and authorization events. If those artefacts are not captured during execution, they cannot be reconstructed reliably later.

For research, benchmark performance is only one input into assessment. A governable workflow must be evaluated for output quality and for reconstructable process evidence, including the set-tings of planning depth, retrieval scope, verification strictness, and control constraints. Existing LLM and agent evaluations remain fragmented [23, p. 9; 25, p. 2], while defence deployment also requires fit with review chains, doctrine, and command-authority structures [10, p. 3; 11, p. 39].

5.3. Limitations

This study does not empirically validate the framework. It specifies governance conditions for agentic defence workflows, but does not establish that compliant systems produce more accurate assessments or better decisions. The next step is a case study or proof-of-concept evaluation comparing workflows with and without the three layers instantiated, measuring both output quality and audit trail completeness. The framework also abstracts from implementation constraints. Context-window limits, model-specific provenance handling, retention rules, and classification constraints may reduce the feasibility of persistent artefact capture or full audit logging. Any deployment would require local gap analysis before compliance conditions can be treated as achievable. Construct validity is a further limitation. Bounded autonomy and the transparency–traceability–reconstructability–auditability distinction need to be tested against practitioner understanding and operational doctrine. Future work should examine whether analysts, system designers, and auditors interpret these constructs consistently.

Adversarial risks are summarized but not resolved. Tool-using and retrieval-dependent systems remain exposed to prompt injection, poisoned evidence pools, and provenance manipulation [26, p. 1–30, p. 1]. The framework addresses governance under normal operating conditions; adversarial assurance remains a separate research task.

6. Conclusion

This paper argued that reliability and accountability in agentic analytical workflows are archi-tectural properties. For defence decision support, the relevant governance question is not only what an LLM can generate, but how the workflow plans, retrieves, verifies, synthesizes, logs, and authorizes its outputs.

The contribution is an audit-first governance framework that operationalizes the I→E→R model's externalization insight through explicit task decomposition, evidence grounding with provenance, and bounded control with logging and authorization. Five roles, four design vari-ables, and bounded autonomy make the process inspectable and constrain where unsupported claims can enter. The conceptual evaluation found the framework coherent, complete with respect to the derived requirements, and plausible against NATO, EU AI Act, and UK defence assurance guidance. It does not prove empirical superiority. It specifies the conditions under which agentic analytical workflows can remain institutionally defensible and prepares the design for proof-of-concept testing.

Acknowledgements. The work was supported by the Ministry of Defence of the Czech Republic, "Long Term Organi-zation Development Plan" – Conduct of Land Operations of the Faculty of Military Leadership, Brno, University of Defence, Czech Republic (Project No. DZRO-FVL22-LANDOPS)

References

1. SYPHERD, Chris and BELLE, Vaishak. Practical Considerations for Agentic LLM Sys-tems. arXiv preprint arXiv:2412.04093. Online. 2024. DOI 10.48550/arXiv.2412.04093.
2. LI, Xinyi, WANG, Sai, ZENG, Siqi, WU, Yu and YANG, Yi. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth. Online. October 2024. Vol. 1, p. 9. DOI 10.1007/s44336-024-00009-2. Review article; published 08 October 2024; article number 9
3. JI, Ziwei, LEE, Nayeon, FRIESKE, Rita, YU, Tiezheng, SU, Dan, XU, Yan, ISHII, Etsuko, BANG, Yejin, CHEN, Delong, DAI, Wenliang, CHAN, Ho Shu, MADOTTO, Andrea and FUNG, Pascale. Survey of Hallucination in Natural Language Generation. ACM Comput-ing Surveys. Online. 2022. DOI 10.1145/3571730.
4. BANG, Yejin, CAHYAWIJAYA, Samuel, LEE, Nayeon, DAI, Wenliang, SU, Dan, WILIE, Bryan, LOVENIA, Holy, JI, Ziwei, YU, Tiezheng, CHUNG, Willy, DO, Quyet V., XU, Yan and FUNG, Pascale. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In : PARK, Jong C., ARASE, Yuki, HU, Baotian, LU, Wei, WIJAYA, Derry, PURWARIANTI, Ayu and KRISNADHI, Adila Alfa (eds..), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Com-putational Linguistics (Volume 1: Long Papers). Online. Nusa Dua, Bali : Association for Computational Linguistics, November 2023. p. 675–718. DOI 10.18653/v1/2023.ijcnlp-main.45.
5. YUAN, Tongxin, HE, Zhiwei, DONG, Lingzhong, WANG, Yiming, ZHAO, Ruijie, XIA, Tian, XU, Lizhen, ZHOU, Binglin, LI, Fangqi, ZHANG, Zhuosheng, WANG, Rui and LIU, Gongshen. R-Judge: Benchmarking Safety Risk Awareness for LLM Agents. In : AL-ONAIZAN, Yaser, BANSAL, Mohit and CHEN, Yun-Nung (eds..), Findings of the Association for Computational Linguistics: EMNLP 2024. Online. Miami, Florida, USA : Association for Computational Linguistics, November 2024. p. 1467–1490. DOI 10.18653/v1/2024.findings-emnlp.79.
6. ZHAO, Qiwei, LI, Dong, LIU, Yanchi, CHENG, Wei, SUN, Yiyou, OISHI, Mika, OSAKI, Takao, MATSUDA, Katsushi, YAO, Huaxiu, ZHAO, Chen, CHEN, Haifeng and ZHAO, Xujiang. Uncertainty Propagation on LLM Agent. In : CHE, Wanxiang, NABENDE, Joyce, SHUTOVA, Ekaterina and PILEHVAR, Mohammad Taher (eds..), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Online. Vienna, Austria : Association for Computational Linguistics, July 2025. p. 6064–6073. ISBN 979-8-89176-251-0. DOI 10.18653/v1/2025.acl-long.302.
7. PARASURAMAN, Raja and RILEY, Victor. Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors. 1997. Vol. 39, no. 2, p. 230–253. DOI 10.1518/001872097778543886.
8. SKITKA, Linda J., MOSIER, Kathleen and BURDICK, Marilyn. Does Automation Bias Decision-Making?. International Journal of Human-Computer Studies. 1999. Vol. 51, no. 5, p. 991–1006. DOI 10.1006/ijhc.1999.0252.
9. LANGER, Markus, BAUM, Kevin and SCHLICKER, Nadine. Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs. Minds and Machines. Online. 2025. Vol. 35, no. 1, p. 1. DOI 10.1007/ s11023-024-09701-0. Published 05 November 2024; article number 1
10. NORTH ATLANTIC TREATY ORGANIZATION. Summary of NATO's revised Artifi-cial Intelligence (AI) strategy. Online. July 2024. Available from: https://www.nato. int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategyOfficial text, 10 July 2024
11. UK MINISTRY OF DEFENCE. JSP 936: Dependable Artificial Intelligence (AI) in defence (part 1: directive). Online. November 2024. Available from: https://www. gov.uk/government/publications/jsp-936-dependable-artificial-intelligence-ai-in-defence-part-1-directivePublished 13 November 2024
12. ROUSEK, Antonin. Agentic Workflows for Structured Analytic Techniques. In: 20th International Doctoral Conference. 2025. Conference proceedings
13. WU, Qingyun, BANSAL, Gagan, ZHANG, Jieyu, WU, Yiran, LI, Beibin, ZHU, Erkang, JIANG, Li, ZHANG, Xiaoyun, ZHANG, Shaokun, LIU, Jiale, AWADALLAH, Ahmed Hassan, WHITE, Ryen W., BURGER, Doug and WANG, Chi. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv preprint arXiv:2308.08155. Online. 2023. DOI 10.48550/arXiv.2308.08155.
14. LEWIS, Patrick, PEREZ, Ethan, PIKTUS, Aleksandra, PETRONI, Fabio, KARPUKHIN, Vladimir, GOYAL, Naman, KÜTTLER, Heinrich, LEWIS, Mike, YIH, Wen-tau, ROCK-TÄSCHEL, Tim, RIEDEL, Sebastian and KIELA, Douwe.

Retrieval-Augmented Genera-tion for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401. Online. 2020. DOI 10.48550/arXiv.2005.11401.
15. KATSIS, Yannis, ROSENTHAL, Sara, FADNIS, Kshitij, GUNASEKARA, Chulaka, LEE, Young-Suk, POPA, Lucian, SHAH, Vraj, ZHU, Huaiyu, CONTRACTOR, Danish and DANILEVSKY, Marina. mt RAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems. Transactions of the Association for Computational Linguistics. Online. 2025. Vol. 13, p. 784–808. DOI 10.1162/tacl.a.19.
16. OUYANG, Jie, PAN, Tingyue, CHENG, Mingyue, YAN, Ruiran, LUO, Yucong, LIN, Jiaying and LIU, Qi. HoH: A Dynamic Benchmark for Evaluating the Impact of Outdated Information on Retrieval-Augmented Generation. In : CHE, Wanxiang, NABENDE, Joyce, SHUTOVA, Ekaterina and PILEHVAR, Mohammad Taher (eds..), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Online. Vienna, Austria : Association for Computational Linguistics, July 2025. p. 6036–6063. ISBN 979-8-89176-251-0. DOI 10.18653/v1/2025.acl-long.301.
17. AUTIO, Chloe, SCHWARTZ, Reva, DUNIETZ, Jesse, JAIN, Shomik, STANLEY, Martin, TABASSI, Elham, HALL, Patrick and ROBERTS, Kamie. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile Online. NIST AI 600-1. Gaithersburg, MD, 2024.
18. EUROPEAN UNION. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Online. 2024. Available from: https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=LEGISSUM:4762484Official Jour-nal of the European Union, 12 July 2024
19. MOREAU, Luc and MISSIER, Paolo. PROV-DM: The PROV Data Model. Online. April 2013. Available from: https://www.w3.org/TR/prov-dm/W3C Recommendation, 30 April 2013
20. MÖKANDER, Jakob, SCHUETT, Jonas, KIRK, Hannah R. and FLORIDI, Luciano. Auditing Large Language Models: A Three-Layered Approach. AI and Ethics. 2024. Vol. 4, no. 4, p. 1085–1115. DOI 10.1007/s43681-023-00289-2.
21. PHERSON, Randolph H. and HEUER, Jr., Richards J. Structured Analytic Techniques for Intelligence Analysis. 3. Thousand Oaks, California : CQ Press, an imprint of SAGE Publications, Inc., 2021. ISBN 9781506368931.
22. A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis Online. technical report. Washington, DC, 2009. Available from: https://www.cia.gov/ resources/csi/static/Tradecraft-Primer-apr09.pdfPrepared by the US Government
23. LASKAR, Md Tahmid Rahman, ALQAHTANI, Sawsan, BARI, M Saiful, RAH-MAN, Mizanur, KHAN, Mohammad Abdullah Matin, KHAN, Haidar, JAHAN, Israt, BHUIYAN, Amran, TAN, Chee Wei, PARVEZ, Md Rizwan, HOQUE, Enamul, JOTY, Shafiq and HUANG, Jimmy Xiangji. A Systematic Survey and Critical Review on Evalu-ating Large Language Models: Challenges, Limitations, and Recommendations. In : AL-ONAIZAN, Yaser, BANSAL, Mohit and CHEN, Yun-Nung (eds..), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Online. Miami, Florida, USA : Association for Computational Linguistics, November 2024. p. 13785–13816. DOI 10.18653/v1/2024.emnlp-main.764.
24. TSAMADOS, Andreas, FLORIDI, Luciano and TADDEO, Mariarosaria. Human control of AI systems: from supervision to teaming. AI and Ethics. Online. May 2025. Vol. 5, p. 1535–1548. DOI 10.1007/s43681-024-00489-4.
25. MOHAMMADI, Mahmoud, LI, Yipeng, LO, Jane and YIP, Wendy. Evaluation and Benchmarking of LLM Agents: A Survey. arXiv preprint arXiv:2507.21504. Online. 2025. DOI 10.48550/arXiv.2507.21504.
26. OWASP GEN AI SECURITY PROJECT. LLM01:2025 Prompt Injection. Online. 2025. Available from: https://genai.owasp.org/llmrisk/llm01-prompt-injection/
27. CHISMON, Dave. Prompt injection is not SQL injection (it may be worse). Online. December 2025. Available from: https://www.ncsc.gov.uk/blog-post/prompt-injection-is-not-sql-injectionBlog post, published 8 December 2025
28. PEREZ, Fábio and RIBEIRO, Ian. Ignore Previous Prompt: Attack Techniques for Language Models. CoRR. Online. 2022. DOI 10.48550/arXiv.2211.09527. Presented at the NeurIPS 2022 Workshop on Machine Learning Safety
29. ZHAN, Qiusi, FANG, Richard, PANCHAL, Henil Shalin and KANG, Daniel. Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents. In : CHIRUZZO, Luis, RITTER, Alan and WANG, Lu (eds..), Findings of the Association for Computational Linguistics: NAACL 2025. Online. Albuquerque, New Mexico: Association for Computational Linguistics, April 2025. p. 7116–7132. ISBN 979-8-89176-195-7. DOI 10.18653/v1/2025.findings-naacl.395.
30. AN, Hengyu, ZHANG, Jinghuai, DU, Tianyu, ZHOU, Chunyi, LI, Qingming, LIN, Tao and JI, Shouling. IPIGuard: A Novel Tool Dependency Graph-Based Defense Against Indirect Prompt Injection in LLM Agents. In: CHRISTODOULOPOULOS, Christos, CHAKRABORTY, Tanmoy, ROSE, Carolyn and PENG, Violet (eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Online. Suzhou, China: Association for Computational Linguistics, November 2025. p. 1023–1039. ISBN 979-8-89176-332-6. DOI 10.18653/v1/2025.emnlp-main.53.