

AI Pipeline Applied to an Adaptive Multistage Decision Support System for Strategic Intelligence

Caio César dos Santos Cruz ^{1,*} , Robson Oliveira de Albuquerque ^{1,2,*} , João José Costa Gondim ¹ , Dino Macedo Amaral ¹  and Fábio Lúcio Lopes de Mendonça ¹ ,

¹ Professional Post-Graduate Program in Electrical Engineering (PPEE), Department of Electrical Engineering (ENE), Faculty of Technology, University of Brasília (UnB), Brasília-DF, zipcode 70910-900, Brazil (caiocscruz@gmail.com, robson@redes.unb.br, gondim@unb.br, dinoamaral@bb.com.br, fabio.mendonca@unb.br)

² Postgraduate in Governance, Technology, and Innovation, Catholic University of Brasília (UCB), Brasília-DF, zipcode 71966-700, Brazil

* Correspondence: caiocscruz@gmail.com, robson@redes.unb.br

Abstract

Large Language Models (LLMs) can automate parts of the intelligence cycle, but hallucinations, context drift, and weak multi-hop reasoning make them unreliable for long context analysis and reporting—especially when data cannot leave local infrastructure. We present a six-stage pipeline modeled on the NATO Intelligence Cycle and guided by the 5W3H framework, in which a supervisory agent coordinates specialized sub-agents for planning, retrieval, analysis, and reporting. The main contribution is the Hierarchical Chain-of-Density (H-CoD): the system builds a structural outline first, then refines density section by section so that long reports stay information-rich without losing coherence. Outputs are cross-checked through Multi-Agent Debate and Self-Consistency voting. On 25 analytical tasks, the pipeline—running open-source models on consumer GPUs—outperformed commercial baselines on factual accuracy, entity density, and reasoning coherence, pointing to architectural design rather than model scale as the more effective lever.

Keywords: Agentic AI; Hierarchical Chain-of-Density; NATO Intelligence Cycle; Multi-Agent Debate; System 2 Reasoning; Green AI; Decision Support Systems.

1. Introduction

Intelligence analysts in Open Source Intelligence (OSINT) and Cyber Threat Intelligence (CTI) are drowning in data. Reports that used to take weeks now demand synthesis across thousands of unstructured documents, and teams have not grown to match. LLMs are an obvious candidate, but two problems get in the way. The first is technical. LLMs hallucinate, and the problem worsens in long contexts where key evidence sits deep in the input [1]. The *lost-in-the-middle* effect [2] shows that larger context windows do not automatically mean better answers, and flat vector retrieval in standard Retrieval-Augmented Generation (RAG) breaks down when queries require chaining facts across sources [3]. The second is operational: in national-security settings, sending data to the cloud is not allowed, so organizations rely on manual analysis.

This paper tackles both issues with a six-stage pipeline built around the NATO Intelligence Cycle [4,14] and the 5W3H framework [13]. Planning and generation are kept separate—a supervisory agent decomposes each task before handing pieces to specialized sub-agents. Our main contribution is the **Hierarchical Chain-of-Density (H-CoD)**, which

Published:

Copyright: © 2026 by the authors.

Submitted to *Eng. Proc.* for possible open access publication under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](#) license.

applies iterative densification of Chain-of-Density (CoD) prompting [7] inside hierarchical outlining of DOME [6]: outline first, compress each section on its own so global structure survives while information per paragraph goes up. The pipeline also uses graph-based retrieval [8] for multi-hop queries, Tree of Thoughts [9], Self-Consistency voting [10], and Multi-Agent Debate [11]. All runs on open-source LLMs and consumer GPUs. In a 25-task evaluation where four commercial LLMs acted as judges, our pipeline scored higher on accuracy, density, and coherence—evidence that architecture matters more than scale, in line with Green AI [12].

2. Proposed Approach

We picked the NATO Intelligence Cycle as backbone [4,14]: Direction, Collection, Processing, Analysis, Dissemination. Adding an Observation stage for input cleanup gives six stages: **Observation, Direction, Collection, Processing, Analysis, Reporting** (Figure 1). Each stage writes to a shared epistemic state object instead of forwarding raw text, so structured knowledge accumulates across the pipeline.

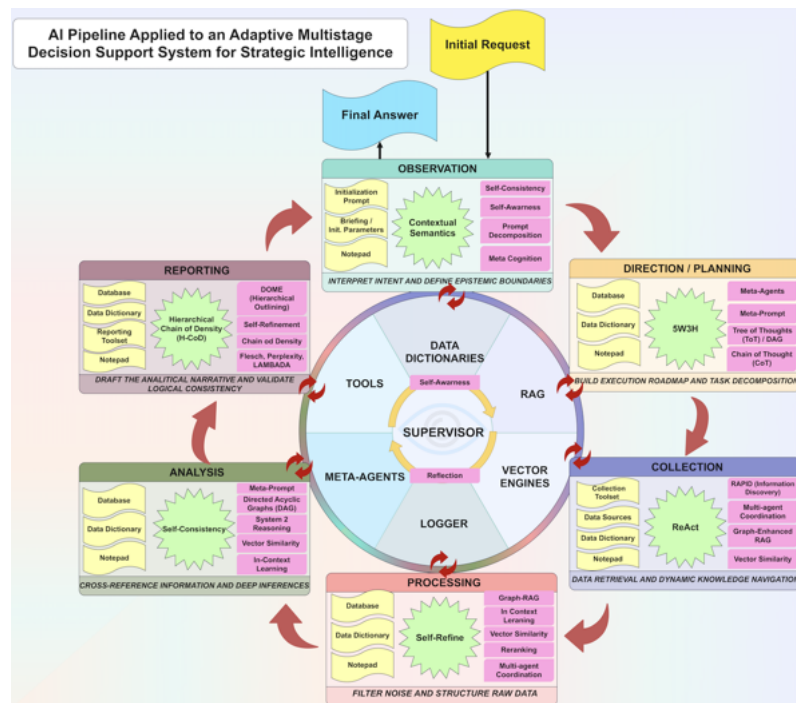


Figure 1. Six-stage adaptive loop with shared meta-cognitive supervision, bidirectional feedback, and persistent epistemic state.

2.1. Pipeline Overview

Observation turns free-form requests into formal requirements. It strips noise and runs a zero-shot complexity classifier. Simple queries get a direct answer; only multi-source problems enter the full pipeline. Output is a JSON contract specifying audience, scope, format, and constraints.

Direction decomposes that contract via 5W3H into sub-questions assigned to retrieval or analysis tasks. **Collection** queries three back-ends: relational (structured metadata), vector (semantic similarity), and graph (entity relationships), and a Crawlers gateway to bring data from OSINT probes. We ended up with this combination because vector-only retrieval simply did not work for queries needing two or more reasoning hops [8].

Processing prepare data for analytics. **Analysis** is where the real reasoning happens: Chain-of-Thought (CoT) planning [16], Self-Refinement loops [15], and In-Context Learning,

under a meta-agentic supervisor that restarts sub-agents when intermediate results do not match quality criteria. **Reporting** produces the final document via H-CoD (Section 2.2). The knowledge base grows with each run, making related queries faster; a hit-rate deprecation policy prunes stale entries.

2.2. Hierarchical Chain-of-Density (H-CoD)

Standard CoD prompting [7] works for short summaries but falls apart on longer reports: densification packs entities into the first paragraphs while later sections drift or repeat. H-CoD splits generation into two levels. A planning agent builds a *Skeleton-of-Thought*—a section-by-section outline with target entity counts and scope constraints, drawing on DOME [6]. Independent sub-agents then run CoD iteratively within each bounded section; densification stays local but the global structure holds. Density is dynamically adjusted based on supporting evidence and adjustable parameters.

Our ML model scores vectorial similarity to catch semantic drift, yielding hallucination rates below 5% on ~20 000-token documents, even from contexts exceeding 10M tokens.

3. Preliminary Evaluation

We challenged our pipeline across 18 complex tasks, including geopolitical risk assessment, supply-chain mapping, threat actor analysis, technology trend forecasting, and regional conflict scenarios — the kind of requests for which no single source or single-step process would suffice — against three commercial LLM models.

The tasks were randomly split into three groups. Each group included one commercial LLM and our model (UNB), competing to produce the best answer. The models that did not participate in that round voted for the best response. The commercial LLMs used were: OpenAI GPT-5.1, Gemini 3, and Claude Sonnet 4.5. Table 1 shows the results.

Match	UNB	Commercial LLMs	Ties
GPT-5.1 x UNB	4	1	1
Gemini-3 x UNB	3	1	2
Claude Sonnet-4.5 x UNB	3	2	1
Total	10	4	4

Table 1. Match on 18 complex queries against three commercial LLMs judged via Multi-Agent Debate.

A limitation is that LLM-as-judge scoring is inherently subjective. Selecting the best answer varies by perspective across multiple rounds; therefore, the relative overall ranking matters more than absolute scores. Human review and automated metrics are planned. All UNB tests used a NVIDIA DGX Spark and two RTX 4060 Ti GPUs, with open-source models no larger than 70B parameters [12].

4. Discussion and Conclusions

Our hypothesis is that the quality problems people associate with LLM-generated analysis—hallucination, drift, shallow reasoning—are artifacts of monolithic, single-pass use without structured methodology or control gates. Break the task into phases, give each right tools, add a supervisor, and output quality goes up with smaller models. "H-CoD integrates two concepts that, while effective independently, fail to achieve an adequate baseline for our use case: CoD’s information packing [7] and DOME’s hierarchical planning [6]. Intelligence reporting needs both to enforce coherence across documents."

Architecturally, the pipeline draws on CoALA’s memory–planning–action separation [18], ReAct’s interleaved reasoning and action [19], and Reflexion’s self-correction [20]. It differs by grounding everything in the NATO Intelligence Cycle [4,14], extending Boyd’s

OODA-loop [21] or some of its more recent adaptations like OADA [5] and recent AI-for-OSINT surveys [22,23]. Graph retrieval [8,17] turned out to be essential, considering simple vector search could not handle entity-chain queries in any of our tests.

Looking ahead, H-CoD combines hierarchical outlining with iterative density refinement so long reports can be both dense and structurally sound. The system runs entirely on local infrastructure with open-source models. Next steps: adversarial robustness testing, CTI feed integration, and further work on the vectorial similarity model for drift detection.

Data Availability Statement: Data and code will be made available upon request to the authors.

Acknowledgments: The authors thank LATITUDE/UnB (Universidade de Brasília), PGFN (Attorney General's Office of the National Treasury), SNAS/DGSUAS (National Secretariat of Social Assistance), TRE-DF (Regional Electoral Court of the Federal District), and the Support Foundation of the Federal District (FAP/DF) for their support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Huang, L.; et al. A Survey on Hallucination in Large Language Models. *ACM Comput. Surv.* **2024**, *57*, 1–38.
2. Liu, N.F.; et al. Lost in the Middle: How Language Models Use Long Contexts. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 157–173.
3. Gao, Y.; et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2312.10997.
4. NATO Standardization Office. *AJP-2.1: Allied Joint Doctrine for Intelligence Procedures*; NATO Standardization Office: Brussels, Belgium, 2016.
5. Costa Gondim, J.J.; et al. A Methodological Approach for Assessing Amplified Reflection Distributed Denial of Service on the Internet of Things. *Sensors* **2016**, *16*, 1855.
6. Gu, Y.; et al. DOME: Dynamic Hierarchical Outlining with Memory-Enhancement for Long-Text Generation. *arXiv* **2024**, arXiv:2405.14739.
7. Adams, G.; et al. From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting. In *Proc. EMNLP NewSumm*, 2023; pp. 68–81.
8. Edge, D.; et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv* **2024**, arXiv:2404.16130.
9. Yao, S.; et al. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Proc. NeurIPS*, 2023.
10. Wang, X.; et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proc. ICLR*, 2023.
11. Du, Y.; et al. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *Proc. ICML*, 2024.
12. Schwartz, R.; et al. Green AI. *Commun. ACM* **2020**, *63*, 54–63.
13. de Melo e Silva, A.; et al. A Methodology to Evaluate Standards and Platforms within Cyber Threat Intelligence. *Future Internet* **2020**, *12*, 108.
14. Lowenthal, M.M. *Intelligence: From Secrets to Policy*, 8th ed.; CQ Press: Washington, DC, USA, 2019.
15. Madaan, A.; et al. Self-Refine: Iterative Refinement with Self-Feedback. In *Proc. NeurIPS*, 2023.
16. Wei, J.; et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proc. NeurIPS*, 2022.
17. Sarmah, B.; et al. HybridRAG: Integrating Knowledge Graphs and Vector Retrieval. In *Proc. ICAIF*, 2024.
18. Summers, T.; Yao, S.; Narasimhan, K.; Griffiths, T. Cognitive Architectures for Language Agents. *Trans. Mach. Learn. Res.* **2024**, Article ID 2024-001.
19. Yao, S.; et al. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proc. ICLR*, 2023.
20. Shinn, N.; et al. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Proc. NeurIPS*, 2023.
21. Johnson, J. Automating the OODA Loop in the Age of Intelligent Machines. *Def. Stud.* **2023**, *23*, 673–697.
22. Hayes, D.R.; et al. A Systematic Review on AI for Open Source Intelligence Applications. *Int. J. Inf. Secur.* **2024**, *23*, 2911–2932.
23. Skopik, F.; et al. On the Application of NLP for Advanced OSINT Analysis in Cyber Defence. In *Proc. ARES*, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.