International Journal of Artificial Intelligence, Data Science, and Machine Learning
Grace Horizon Publication | Volume 7, Issue 2, 112-116, 2026
ISSN: 3050-9262 | https://doi.org/10.63282/3050-9262.IJAIDSML-V7I2P117

Original Article

Reflexion-Based Agentic Content Review: An LLM-as-a-Judge Framework for Lakehouse ECM

Vamshi Krishna Malthummeda,
Independent Researcher, USA.

Received On: 08/03/2026 ***Revised On: 02/04/2026*** ***Accepted On: 09/04/2026*** ***Published On: 20/04/2026***

Abstract *- Enterprise departments such as legal and marketing uses content management systems like SharePoint and Salesforce for collaboration on large volume of marketing campaign and legal contracts content. The content needs to be fact-checked, compliant, relevant, consistent and of high-quality, to ensure all this is a time-consuming, manual and error-prone process. This paper presents a databricks based automated content review framework that leverages Reflexion-agent based agentic architecture where LLMs (large language models) play the role of judges to evaluate, critique and refine the content. Enterprise content ranging from marketing campaign articles to legal contracts are ingested into databricks Lakehouse using scalable, robust and self-healing ingestion pipelines. The proposed Reflexion agent is configured using a powerful LLM, detailed system prompts with clear rubrics to ensure clarity, compliance and brand adherence by generating critiques and self-reflections iteratively to finally achieve the improved outcome without human involvement. Experimental results indicate accurate and automated content review across marketing and legal content repositories. This work highlights the potential of Reflexion agent-based content review systems and establishes Databricks as a platform for intelligent content life cycle management.*

Keywords *- Reflexion Agent, Salesforce, LLM-as-a-Judge, Content Governance, SharePoint, Langchain, Unity Catalog, Lakehouse.*

1. Introduction

Informative and valuable marketing content builds trust and addresses the needs of potential and existing customers. Well-drafted content in legal contracts protects both the parties, precisely outline responsibilities, rights, and obligations, preventing ambiguity. This kind of content is distributed across various document repositories in Microsoft SharePoint and various objects (standard and custom) in Salesforce.

These platforms store below enterprise content:

- Marketing Campaign: Marketing Plans, Strategy Documents, Market Research Reports, Marketing Proposals, Campaign Narratives, Product descriptions, Compliance Disclosures, and Customer-facing Communications.
- Legal Contracts: Employment Contracts, Non-Disclosure Agreements (NDAs), Service Agreements, Loan Agreements, Sales Contracts, Partnership Agreements, License Agreements, Copyright Assignment, Purcha Orders etc.

As enterprise content volume and velocity increase, ensuring quality, consistency and adherence to the evaluation rubric while reviewing the content becomes highly challenging.

Traditional content review process involved triggering workflows on upload of content to the libraries in the SharePoint and objects in Salesforce. As part of workflow process an email notification will be sent to reviewers to review the content uploaded. Content reviewed by humans is time consuming, error-prone, subjective with availability being limited. Some kind of automation and speed can be achieved by employing LLMs and by using one-shot evaluation but the accuracy level, refinement and adherence to evaluation criteria of output is low.

This paper proposes a Reflexion-based agentic Content Review Framework which integrates Reflexion-based agents with an LLM-as-judge paradigm, deployed on a Lakehouse which unifies and governs the content from enterprise content management systems like Microsoft SharePoint, Salesforce etc. The proposed system provides automated, accurate, scalable and auditable reviews of the content.

2. Background and Related Work

2.1. Automated Content Review Using Llms

LLMs have been applied to content summarization, sentiment analysis, and quality assessment. In addition, LLM language understanding and generation capabilities are leveraged by the reflection agents to automate the review of enterprise content. A reflection agent is an AI system that improves its output through a generate-critique-refine loop, where it produces an initial response, a separate "reflector" module critiques it, and the generator then revises the output. The exit conditions for generate-critique-refine loop include fixed number of iterations, token limit, absence of errors etc. This process mimics human introspection to catch errors and enhance quality before finalizing a task [[5]].

The limitations of this approach are:

- Susceptibility to repeat the same mistakes in reviewing the content of future tasks due to unavailability of long-term memory
- Improvement of the response is limited by the self-evaluation capabilities of the underlying LLM.
- The agent might get stuck in a loop, repeatedly reflecting on minor variations of the same flawed approach without fundamentally changing its strategy or finding the correct solution.
- Tendency to hallucinate
- The review of the content and the evaluation of the review is performed by same LLM which introduces bias in the final response.

2.2. Reflexion Agent

A Reflexion agent is an advanced AI system that improves its performance by using a cycle of generation, evaluation, and self-reflection, storing insights in an episodic memory to learn from past mistakes and refine future outputs, making it more effective for complex tasks like coding or reasoning without constant model retraining. It critiques its own work, incorporating external feedback or tool use for more grounded improvement, moving beyond simple trial-and-error to learn strategically. The application of the Reflexion agent for enterprise content review is currently very minimal [[2]].

2.3. LLM Acting As a Judge Paradigm

LLM as a judge paradigm separates content review and its evaluation. Here the content is evaluated using fixed, rubric-driven prompts, produces structured scores and decisions to achieve evaluation consistency, reduce bias and support auditability and traceability. But its adoption for enterprise content review evaluation is negligible [[3]].

2.4. Data Lakehouse

A data Lakehouse is mainly used to provide a unified, cost-effective platform for all data analytics, combining the flexibility of data lakes (storing all data types cheaply) with the performance and governance of data warehouses (structured data management). It is widely adopted for workloads like Business Intelligence (BI), machine learning (ML), rarely used for agentic AI workflows and content governance [[4]].

3. Scope of the Paper

This paper discusses a framework where Reflexion agents are integrated with an LLM-as-a-judge evaluation paradigm deployed on a Lakehouse architecture. The scope includes automated review of the textual content sourced from enterprise content management systems like Microsoft SharePoint and Salesforce.

This paper addresses the end-to-end content review lifecycle starting with ingestion of enterprise content along with associated metadata and applicable domain specific rubric prompt using the ingestion pipeline, followed by the content critique, independent evaluation, auditable decision logging and Reflexion-guided revision.

The focus of the paper is to achieve consistent, repeatable and governance-ready content review at enterprise scale by the orchestration of functionally distinct cooperative agents in a distributed fashion.

4. Contribution of the Paper

Below are the novel and main contributions of the paper:

- Novel application of Reflexion-based agents coupled with separate LLM evaluation to automate the review of content at scale.
- Novel system design in which a separate LLM acts judge to evaluate the content reviewed by reviewer agent and Reflexion-agents using domain specific rubric prompts.
- Novel use of the Databricks Lakehouse as a persistent substrate for agent memory, reflection traces, evaluation scores, and versioned content, transforming the Lakehouse into an active component of agentic reasoning and enterprise AI governance
- The framework's ability to evaluate content belonging to various domains by using domain specific rubric prompts.
- The paper provides empirical evidence of improved content quality, reduced manual effort and scalable performance

5. System Architecture

Following are the components of the proposed system:

5.1. Enterprise content ingestion pipeline

The ingestion pipeline is built on Lakehouse to extract data from enterprise content management systems like SharePoint and Salesforce on a predefined schedule.

5.1.1. Below are the steps to extract data from SharePoint and load into Lakehouse

- Register an application in Azure AD with necessary permissions to access the target SharePoint document library via direct oData endpoints and authenticate using OAuth 2.0 client credentials[[6]].
- Lakehouse batch job makes oData calls to the specific SharePoint document library by specifying filter, select, top and skip oData query parameters to extract and save the data into volumes in raw JSON format.
- The raw data is subjected to cleansing, transformation, enrichment and saved into the Delta lake tables.

5.1.2. Below are the steps to extract data from Salesforce and load into Lakehouse

- Authenticate using OAuth 2.0 client credentials with the Salesforce App.
- Salesforce object data is ingested into Databricks Lakehouse Ingestion job by querying standard /services/data/vXX.X/query/ salesforce REST API endpoint by passing SOQL queries as URL parameters, with spaces being replaced by

the + sign (or %20) (If the record volume is high, then we need to use Salesforce BULK API)[[7]].
- The extracted data is saved into volumes in raw JSON format.
- The raw data is subjected to cleansing, transformation, enrichment and saved into the Delta lake tables.

5.2. Lakehouse as Unified Content Repository
- Provides unified cheap object storage for structured textual content
- Provides Versioning and historical audit trail capabilities
- The Lakehouse acts as a scalable, persistent storage system for the agent's experiences, observations, and generated reflections, going beyond the limited context window of an individual LLM.
- The output from the reviewer agent and the critiques/scores from the LLM judge are systematically stored in Lakehouse. This provides a robust dataset for tracking performance metrics, identifying biases, and continuously improving the agents over time.

5.3. Reflexive Agentic Review Pipeline
The review pipeline consists of three primary agents:
- Reviewer Agent: Performs initial critique of content based on clarity, tone, redundancy, and alignment with the objectives.
- Reflexion Agent: Analyzes reviewer feedback, reasons over deficiencies, and iteratively refines the critique or recommendations.
- Judge Agent (LLM-as-a-Judge): Independently evaluates the content using fixed, rubric-driven prompts and produces structured scores and decisions.
- The loop continues until content meets predefined quality thresholds or a maximum iteration count is reached.

5.4. LLM-as-a-Judge Evaluation Framework
5.4.1. Rubric-Based Prompt Design
The Judge agent is constrained by a stable evaluation rubric, including:
- Clarity (1–5)
- Brand alignment (1–5)
- Compliance risk (Low/Medium/High)
- Tone consistency
- Redundancy detection

5.4.2. Structured Output and Decision Making
The Judge produces structured output in machine-readable formats, enabling:
- Automated approval/rejection
- Reflexion-triggered iterations
- Governance dashboards and audits

5.4.3. Bias Reduction and Auditability
By separating judgment from generation and maintaining fixed prompts across iterations, the framework improves evaluation consistency and supports enterprise audit requirements.

5.5. Reflexive Agentic Content Review Flow
- Ingest enterprise content from SharePoint and Salesforce into a governed Databricks Lakehouse.
- Critique content using an agentic reviewer enriched with domain and compliance context.
- Apply Reflexion-based self-analysis with persistent memory to guide iterative improvement.
- Evaluate outputs using an independent LLM-as-a-Judge with fixed, rubric-driven criteria.
- Iterate review and revision until enterprise quality thresholds are met and auditable approval is achieved.

Below is the process flow diagram:

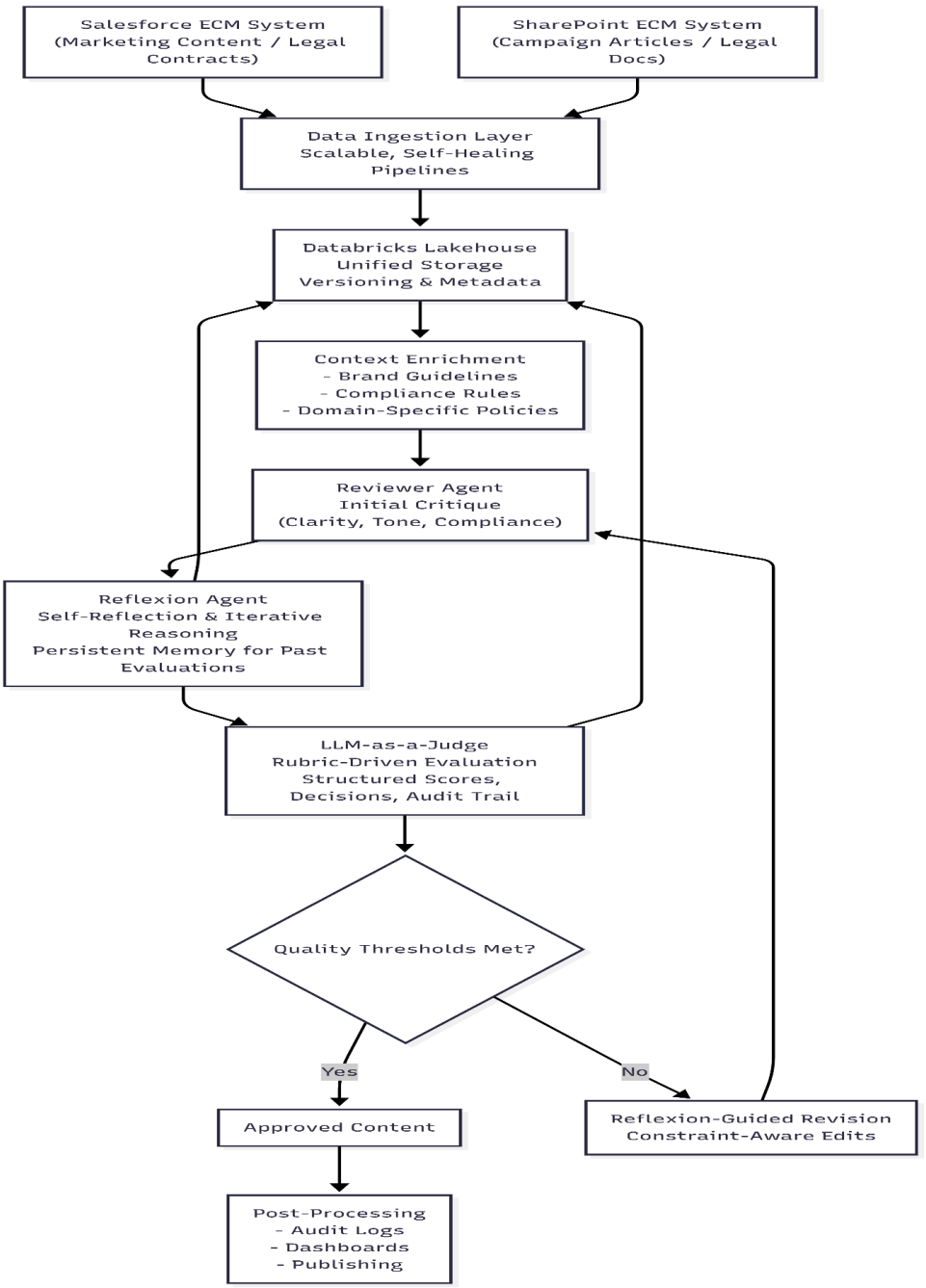

Fig 1: Reflexion-Agent Content Review on Databricks Lakehouse

6. Results and Discussion

To contextualize the effectiveness of reflexive agentic architectures, we reference empirical results reported in the *Multi-Agent Reflexion (MAR)* study[[1]], which evaluates the impact of iterative self-reflection and multi-agent reasoning on question-answering performance. The study reports Exact Match (EM) accuracy on the HotPotQA benchmark using GPT-3.5, comparing baseline and reflexive agent configurations.

The baseline ReAct agent achieves an EM accuracy of 32%, reflecting the limitations of single-pass reasoning and tool use in complex multi-hop question answering tasks. Incorporating single-agent Reflexion significantly improves performance to 44% EM, demonstrating that iterative self-critique and learning from prior failures materially enhance reasoning accuracy. The proposed Multi-Agent Reflexion (MAR) framework further improves EM accuracy to 47%, indicating that collaboration among multiple reflexive agents yields additional gains beyond those achievable with a single reflexive agent.

These results suggest that the performance improvements arise not merely from additional inference steps, but from the structured integration of reflection, memory, and evaluation across agents. The incremental improvement achieved by MAR over single-agent Reflexion highlights the value of diversity in reasoning strategies and

shared reflection memory. The authors also note that the Exact Match metric may underestimate true reasoning capability, as semantically correct answers that differ lexically from ground truth are penalized under strict EM evaluation.

The findings from the MAR study provide strong empirical motivation for the framework proposed in this paper. While the MAR work focuses on question-answering tasks, the demonstrated benefits of reflexive and multi-agent evaluation directly inform our design of a Reflexion-agent-based content review system with an independent LLM-as-a-Judge. In particular, the separation of critique, reflection, and judgment, combined with iterative improvement cycles, aligns with the observed performance gains reported in MAR and supports the applicability of reflexive agentic architectures to enterprise content governance scenarios.

7. Conclusion and Future Work

This paper presents a novel Reflexive Agentic Content Review Framework for enterprise marketing systems, integrating SharePoint and Salesforce content within a Lakehouse architecture and leveraging LLM-as-a-Judge evaluation. The approach advances automated content governance by introducing iterative self-reflection, independent judgment, and scalable enterprise deployment. Future work includes multi-judge ensembles, human-in-the-loop validation, and extension to additional enterprise content systems.

References

[1] Ozer, O., Wu, G., Wang, Y., Dosti, D., Zhang, H., & De La Rue, V. (2025). MAR: Multi-Agent Reflexion Improves Reasoning Abilities in LLMs. *arXiv preprint arXiv:2512.20845*.

[2] Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, *36*, 8634-8652.

[3] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, *36*, 46595-46623.

[4] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of CIDR* (Vol. 8, p. 28).

[5] https://blog.langchain.com/reflection-agents/

[6] Registering the Azure App for SharePoint Online

[7] Vattam, L. (2022). Salesforce REST API in Action: A Practical and Research-Based Exploration of Integration Solutions. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(2), 36-43.