Towards Explainable AI in Advertising Compliance: A Logic-Driven Two-Stage Multi-Agent System – Extended Abstract

Makoto Takamatsu[1], **Modong Tan**[1], **Ryotaro Sugiura**[1], **Mioko Hirose**[2], **Naoya Takizawa**[2]

[1]IBM Japan Systems Engineering Co.,Ltd. (ISE)

[2]IBM Japan Ltd.

Makoto.Takamatsu@ibm.com

Abstract

This study addresses the challenges of transparency and explainability in AI-driven advertising compliance review by proposing an autonomous AI agent that integrates Logic of Thought (LoT) to clarify logical propositions with Reflection Prompting (RP) for self-correction mechanisms. The proposed agent identifies legal violations in the advertising content and expresses them as logical formulae, enabling visualization of reasoning processes previously unattainable in conventional models. We evaluated the agent using 100 experimental cases based on administrative guidance and court decisions related to Japan's Act against Unjustifiable Premiums and Misleading Representations and the Unfair Competition Prevention Act. The results showed significant performance improvements, with accuracy increasing from 72% to 90% in identifying the relevant laws and number of articles, and the F1 score improving from 60% to 85%. The results of this study suggest the potential for flexible adaptation to various legal frameworks across different countries, indicating applicability beyond advertising reviews for general products, financial products, and pharmaceuticals. It also highlights the possibility of application to other tasks requiring explainable and detailed legal compliance assessments.

1 Introduction

In the advertising industry, the importance of advertisement review has grown significantly due to digital transformation and increasing information volume. Inappropriate advertising expressions mislead consumers, damage corporate credibility, and pose legal risks, necessitating efficient and transparent methods for regulatory compliance. Traditional manual review approaches face challenges such as time and cost constraints, as well as risks of subjective bias and oversight. Consequently, there is growing interest in automating advertisement review using artificial intelligence (AI). While natural language processing (NLP) and machine learning technologies have shown promise in detecting regulatory violations, current models face two major challenges:

1. **Insufficient Regulatory Application Accuracy**: Existing models frequently make errors in identifying violations and applying relevant regulations.
2. **Lack of Reasoning Transparency**: Models' inability to adequately explain their decision-making process undermines trust in their results. [Arrieta and others, 2020].

To address these challenges, we propose an autonomous AI agent integrating **Logic of Thought (LoT)**[Liu and others, 2024] and **Reflection Prompting (RP)**[Li and others, 2023]. LoT enhances transparency by explicitly expressing AI reasoning processes as logical propositions, while RP provides mechanisms for AI to self-evaluate and correct reasoning errors. Our research focuses on improving reasoning transparency, accuracy, and cross-domain applicability in regulated fields like financial and pharmaceutical advertising.

2 Method

This study proposes an automated system for verifying legal compliance in advertising content. The system employs a two-stage verification process based on Large Language Models (LLMs) to efficiently and accurately identify and evaluate potential legal violations.

System Overview: As shown in Fig.1, the system has two stages: law-specific agents (Stage 1) and logic-based risk assessment (Stage 2).

Stage 1. Identify and Organize Violations: In the first stage, law-specific agents analyze advertising content and detect expressions that violate each respective law. Each agent stores the full text of its respective law as a key-value cache and loads it during detection. This mechanism supports not only Japanese statutes but also various legal systems of different countries in multiple languages recognized by the LLM. An integration agent aggregates the outputs from individual agents to assess the relationships and severity of violations spanning multiple laws. This produces a comprehensive output detail-

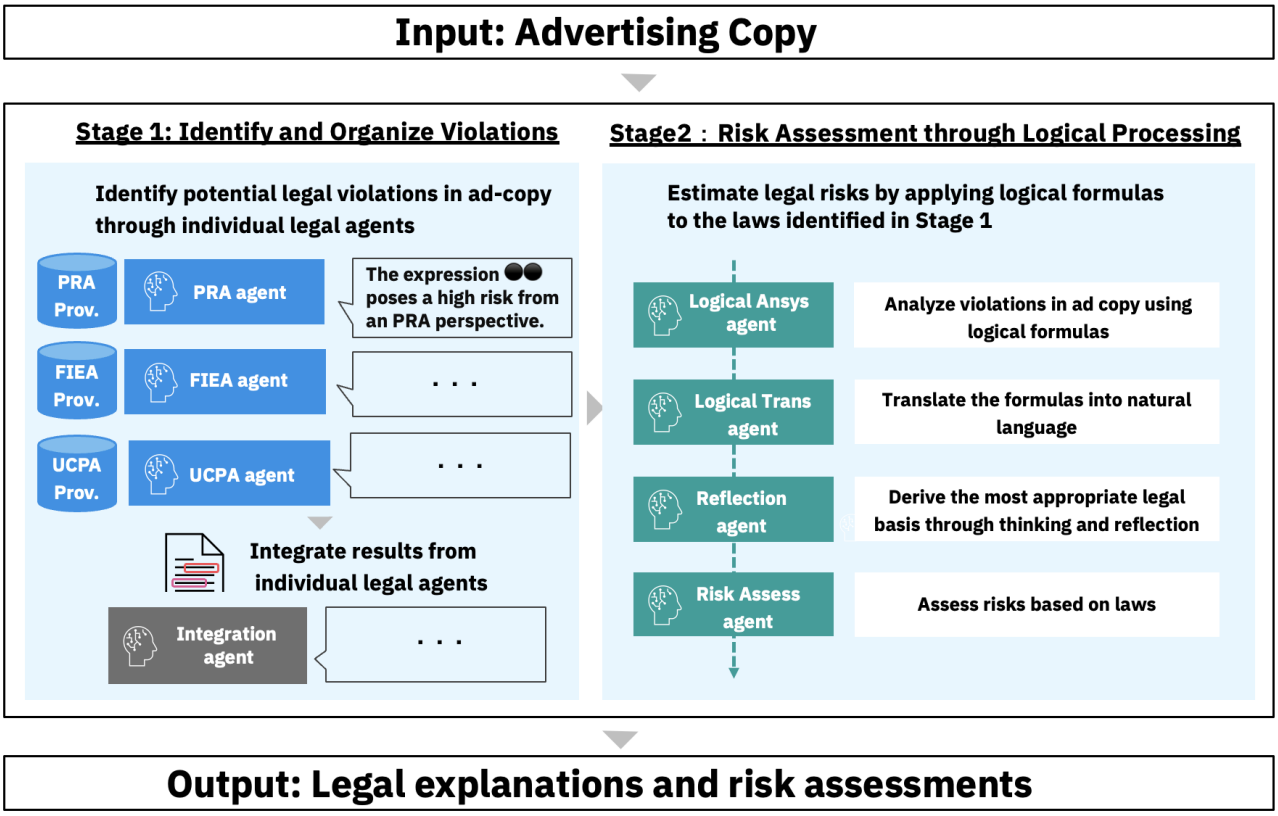

Figure 1: In the first stage, specialized LLM agents for each law independently identify potentially high-risk sections within the input text. Specifically, the PRA (Premiums and Representations Act), FIEA (Financial Instruments and Exchange Act), and UCPA (Unfair Competition Prevention Act) agents store their respective statutes in key-value caches. During the detection process, each agent loads its relevant cache. Subsequently, an integration agent consolidates the outputs from these legal agents. In the second stage, four specialized agents—logical analysis, logical transformation, reflection, and risk evaluation—collaborate to apply logical analysis to the identified risks and provide a comprehensive legal risk assessment. This autonomous multi-agent approach enables efficient and accurate evaluation of advertising text for legal compliance.

ing violation locations, relevant laws, and severity assessments.

Stage 2. Risk Assessment and Legal Analysis: The second stage conducts detailed risk assessment and identifies specific applicable laws and legal codes. A logical analysis agent examines the logical structure of violations using formal logic to generate proposition lists and logical formulae. A logic translation agent converts these into natural language for clarity. A reflection agent then validates the analysis results, evaluating and correcting potential errors in its reasoning. Finally, a risk assessment agent classifies risk levels (high/medium/low) with supporting rationale.

3 Result and Discussion

The system was evaluated using 100 test cases derived from administrative guidance and court decisions related to Japan's Act against Misleading Representations and Unfair Competition Prevention Act. Comparing the complete two-stage system against Stage 1 alone, we observed significant performance improvements in identifying advertising violations and relevant legal codes. The accuracy increased from 72% to 90%, while the F1 score improved from 60% to 85%. These results, achieved using Mistral Large 2 and Llama3.3-70b models, demonstrate that our proposed system effectively enhances both efficiency and transparency in advertising compliance verification.

References

[Arrieta and others, 2020] Alejandro Barredo Arrieta and others. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 2020.

[Li and others, 2023] Ming Li and others. Reflection-tuning: Recycling data for better instruction-tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023.

[Liu and others, 2024] Tongxuan Liu and others. Logic-of-thought: Injecting logic into contexts for full reasoning in large language models. *arXiv preprint arXiv:1904.07734*, 2024.