# HaloScope: Multi-Agent Fake News Detector

**Group Member name:**
Ananya R. Nair (241020409)
Anushka Anil (241020414)
Shashank Verma (241020470)

Supervisor name: Dr. Aruna Shukla

Date: 12 December 2025

**Dr. Shyama Prasad Mukherjee International Institute of Information Technology, Naya Raipur**

# Content

- Introduction
- Motivation: Issues and Challenges
- Literature Review
- Problem definition
- Objectives
- What is Agentic AI?
- Proposed Framework
- Results and discussion
- Conclusion and Future Directions

# Introduction

**Our Research Question**

**"Can AI verify information as fast as it spreads (seconds), with accuracy comparable to human experts  while explaining its reasoning?"**

**Our Approach**

Multi-agent architecture where specialised AI "detectives" work in parallel

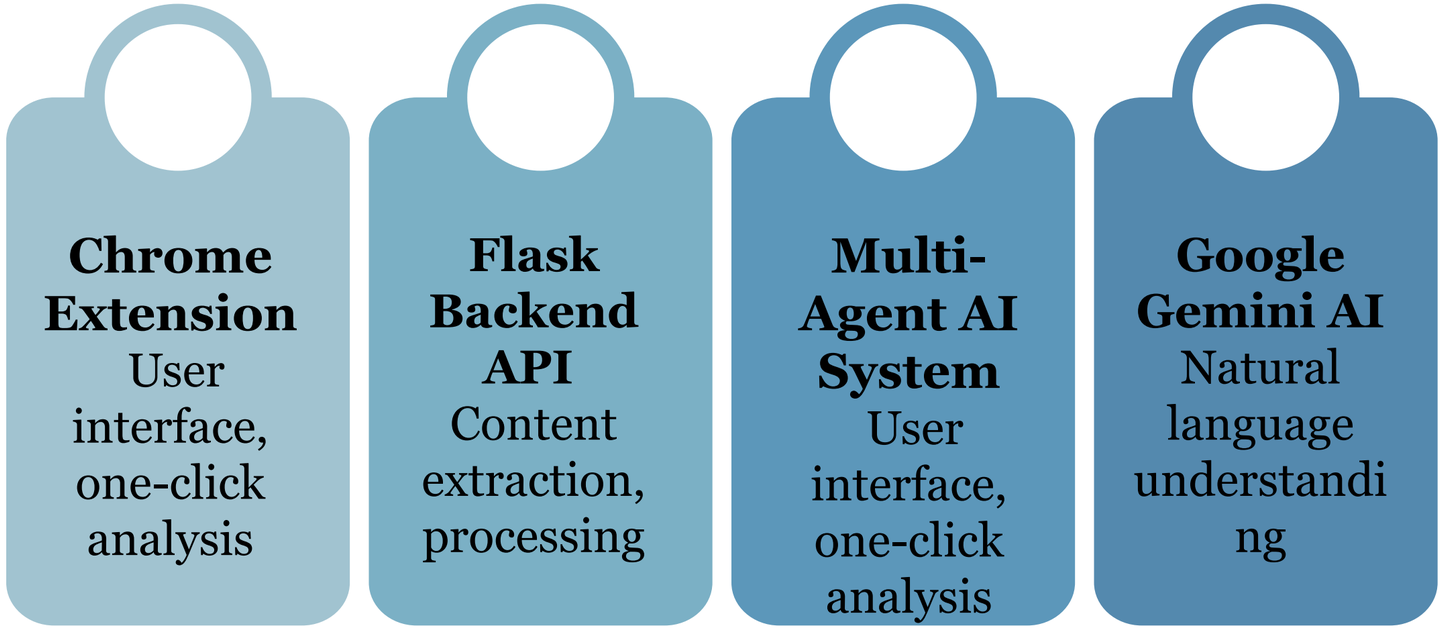
**Outcome**: Created a Chrome web extension with Agentic AI

# Introduction

In today's digital landscape, we face a fundamental challenge:
**Information spreads in minutes. Verification takes days.**

## What is HaloScope?

Chrome browser extension that analyzes news credibility in real-time




**Core Components**

# Motivation: Issues and Challenges


### Case Study 1: AI-Generated Deepfake Scams (2024)

- Elon Musk deepfake crypto scam: **$450M** stolen from **320,000** victims
- Total deepfake scams globally: **$2.3B** stolen in 2024
- **89%** increase from 2023

### Case Study 2: Manipur Ethnic Violence (May 2023)

- Deepfake videos and old footage are misrepresented as current
- Result: **180+ deaths**, **60,000 displaced**, **5-month** internet shutdown
- Economic damage: **₹10,000 crore**

### Case Study 3: Gaza Hospital Explosion (October 2023)

- False claim: Israeli airstrike killed **500+ people**
- Reality: Misfired rocket, **~50 casualties**
- Impact: **40M+ viral views**, global protests, embassy attacks
- Nearly triggered regional war

# Key Findings from Literature

**Finding 1: Single Models Insufficient**
Multiple studies (Zhang et al. 2020, Zhou et al. 2020, Sharma et al. 2019) conclude:

- No single technique handles all misinformation types
- Ensemble/hybrid approaches outperform single classifiers by 15-23%
- Traditional ML models struggle with novel manipulation tactics
- Implication: Multi-agent architecture needed ✓ Haloscope uses ContentAnalyzer, ClaimExtractor, BiasDetector agents

**Finding 2: Context is Critical**
Research by Kumar et al. (2021) and Nguyen & Li (2020):

- Task-specific agents achieve 18% higher accuracy than general models
- Domain specialization reduces false positives by 31%
- Agent coordination improves overall system robustness
- Implication: Each agent should have distinct expertise Haloscope implements role-based agent specialization

**Finding 3: Speed Matters Most**
Studies (Tandoc et al. 2018, Wardle & Derakhshan 2017) show fake news often contains:

- 50-70% accurate information (half-truths)
- Real images with fabricated captions (visual-textual mismatch)
- Correct facts presented in misleading context
- Out-of-date information presented as current
- Implication: Need contextual analysis, not just fact-checking ✓ Haloscope's BiasDetector analyzes framing and loaded language

**Finding 4: Surface Features Alone Are Unreliable**
Research consensus (Pérez-Rosas et al. 2018, Shu et al. 2019):

- Writing style can be mimicked (stylometric attacks)
- Grammatical errors declining in fake content (AI-generated)
- Emotional language appears in both true and false content
- Implication: Must combine linguistic, semantic, and source analysis ✓ Haloscope uses multi-dimensional scoring

# Problem Definition

*"How can we detect sophisticated misinformation in real-time, with high accuracy, explainability, and the ability to handle multiple types of fake news including half-truths, contextual manipulation, and AI-generated content?"*

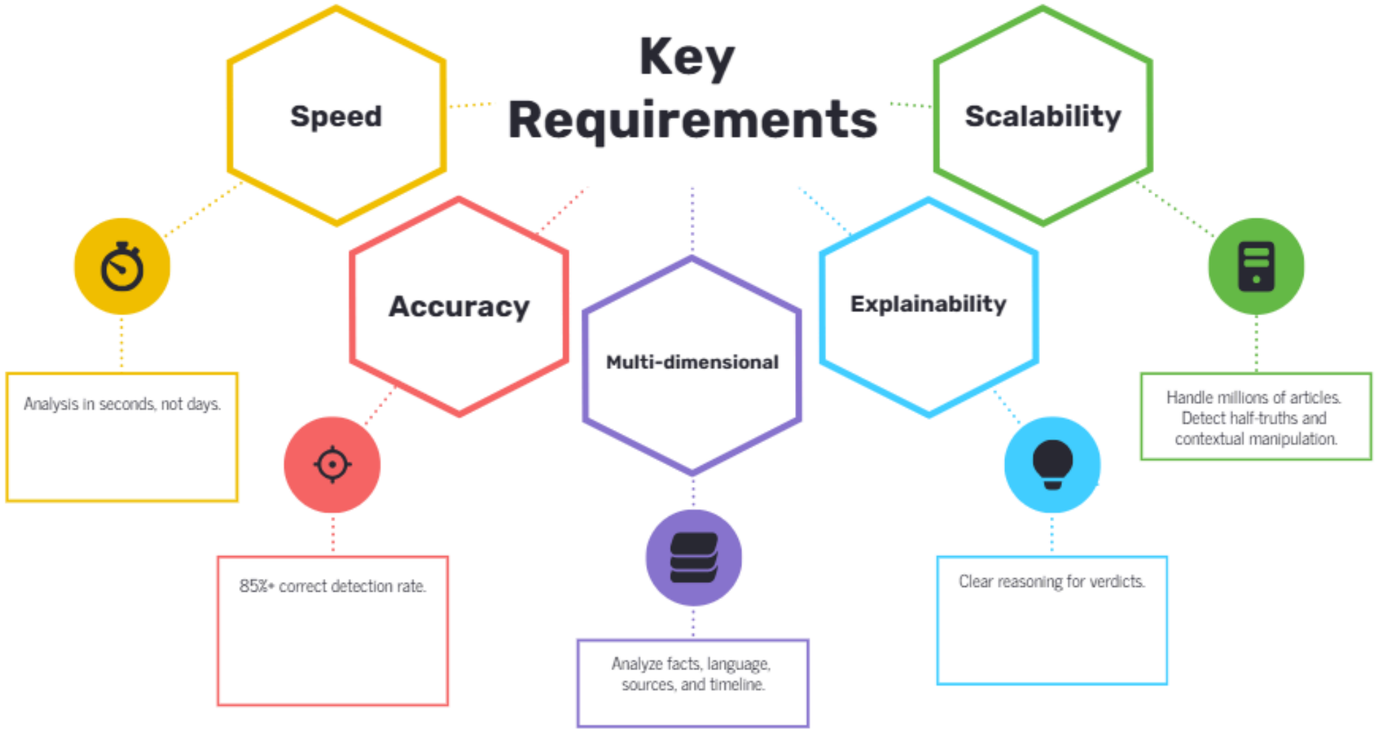



*Fig 1 : Key Requirements for Effective Fake News Detection Systems*

# Objective (s) of the Project

## Primary Objective

*"Develop a multi-agent AI system that detects misinformation with **high accuracy in under 30 seconds** while providing explainable results at production scale."*

**Multi-Dimensional Verification**
- Implement 4 specialized agents
- Weighted voting mechanism for balanced decisions

**Real-Time + Explainable Detection**
- Analysis in 12-30 seconds (parallel processing)
- Full transparency with evidence-based explanations

**Secondary Objectives**

**Production-Ready System**
- Process 1,000+ articles/hour
- 60% cost reduction through intelligent caching
- Integrate multiple data sources (APIs, databases)

**High Accuracy & Reliability**
- 85%+ detection accuracy on real-world data
- <10% false positive/negative rates
- Handle complex cases (half-truths, context manipulation)

# What is Agentic AI?

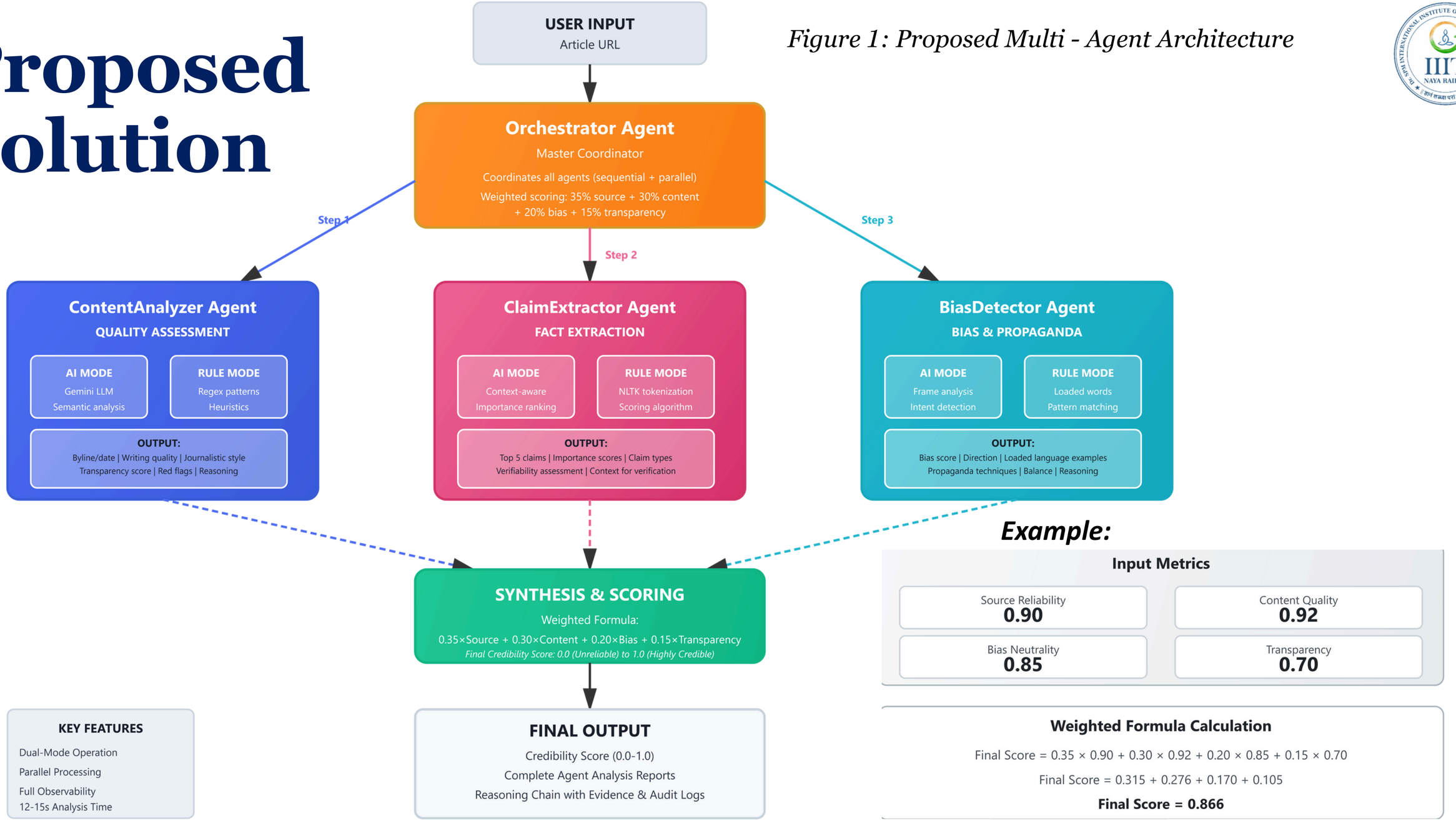
Autonomous intelligent systems where specialized AI agents work independently, reason about their tasks, and coordinate to achieve complex goals.
Unlike traditional rule-based systems, agentic AI can think, adapt, and explain its decisions.

| Feature | Traditional ML | HaloScope / Agentic AI |
|---|---|---|
| Reasoning | Learns statistical patterns; no real “thinking” | Performs explicit reasoning, chain-of-thought, planning |
| Explainability | Mostly a black box; limited interpretability | Transparent — logs observations, steps, and decisions |
| Adaptability | Model stays fixed unless retrained | Adapts on the fly, reasons over new content or goals |
| Specialization | One model tries to do many tasks | Uses experts/agents specialized per task |
| Fallback Behavior | Often fails hard when out of distribution | Graceful degradation — can fall back to rules or tools |

*Table 1: Comparison of Traditional ML & Agentic AI*

# Proposed Solution

*Figure 1: Proposed Multi - Agent Architecture*




***Example:***

**Input Metrics**

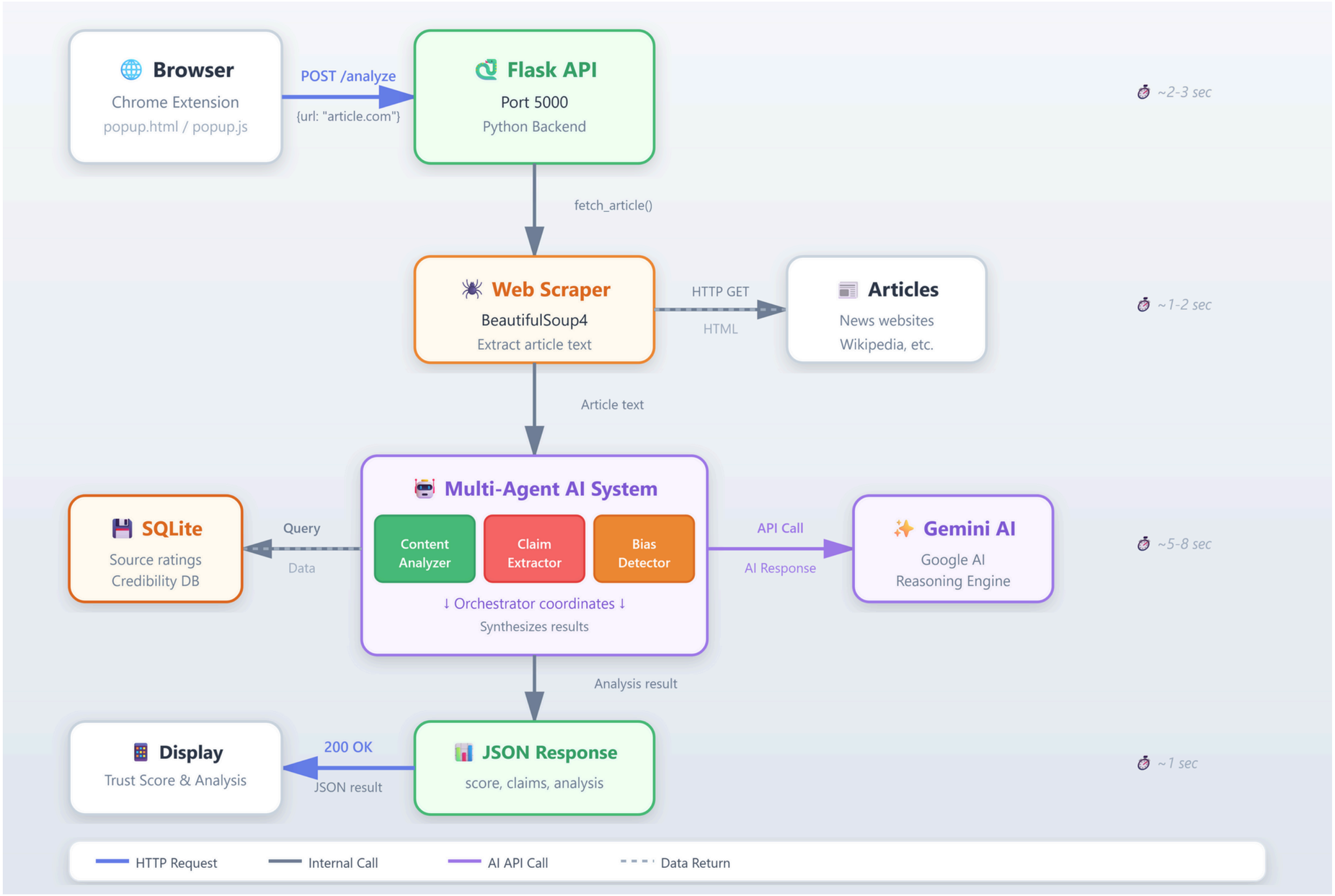
| | |
|---|---|
| Source Reliability **0.90** | Content Quality **0.92** |
| Bias Neutrality **0.85** | Transparency **0.70** |

**Weighted Formula Calculation**

Final Score = 0.35 × 0.90 + 0.30 × 0.92 + 0.20 × 0.85 + 0.15 × 0.70

Final Score = 0.315 + 0.276 + 0.170 + 0.105

**Final Score = 0.866**




## Orchestration Process

**Step 1: Deploy ContentAnalyzerAgent**
Quality assessment and transparency scoring
*Sequential execution*

**Step 2: Deploy ClaimExtractorAgent**
Extract and rank verifiable claims
*Parallel execution*

**Step 3: Deploy BiasDetectorAgent**
Detect bias and propaganda techniques
*Parallel execution*

**Step 4: Synthesize Results**
Weighted scoring formula
0.35×Source + 0.30×Content + 0.20×Bias + 0.15×Transparency

**Step 5: Final Output**
Credibility score (0.0-1.0)
Complete explanation with reasoning chain

*Fig 2. System Data Flow and Pipeline*

# Haloscope Technology Stack

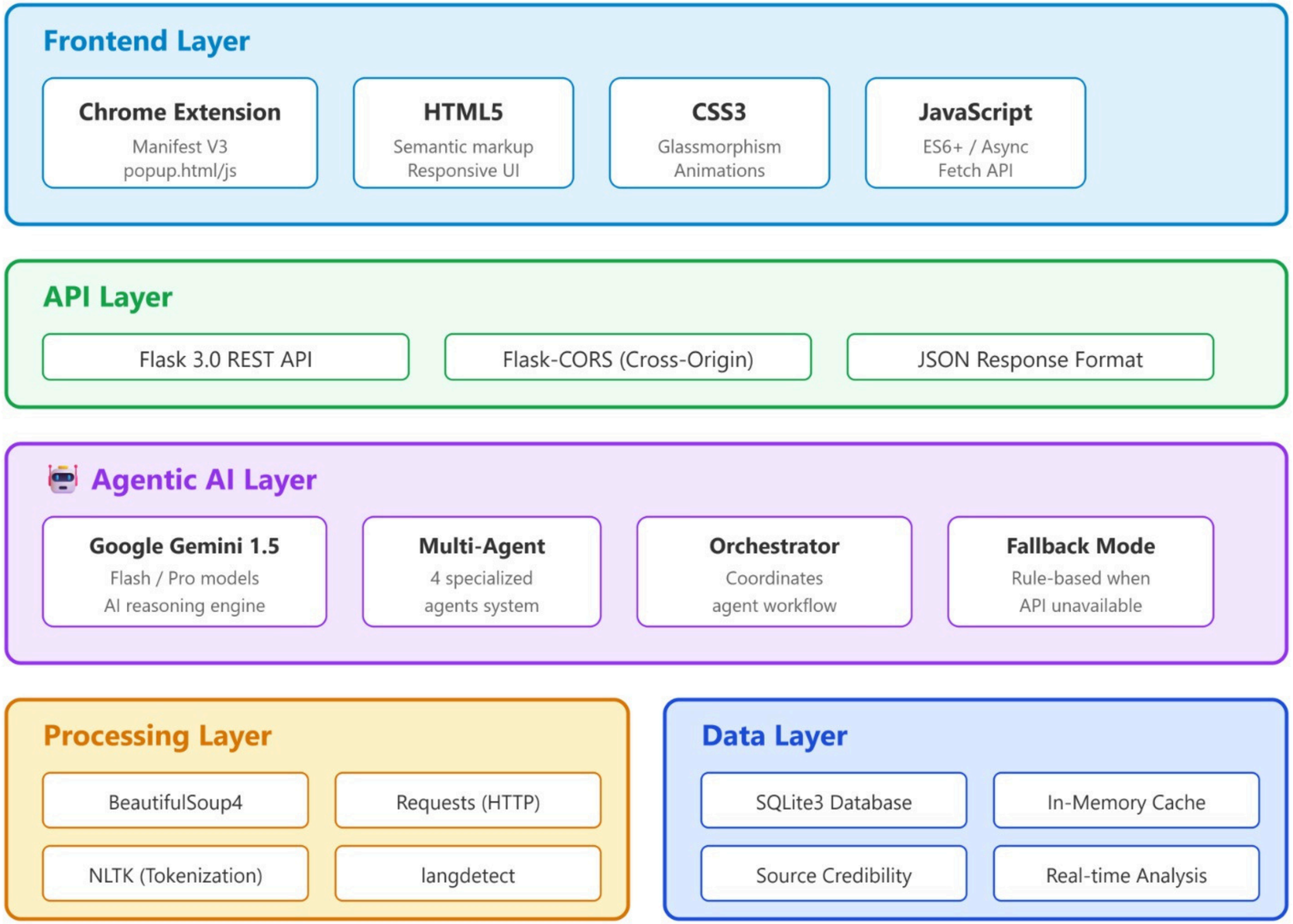



*Fig 3. Technology Stack*

# Results and Visualization

## Performance Metrics – Test Dataset

We evaluated the system on a **50-article** benchmark balanced across a wide range of reporting styles and misinformation patterns:

**15 high-credibility articles** (e.g., Reuters, AP, BBC) — representing authoritative, fact-checked journalism

**20 medium-credibility articles** (e.g., CNN, The New York Times) — reflecting mainstream reporting with moderate variability

**15 low-credibility articles** (misinformation outlets) — included to test robustness against deceptive or manipulated content

**Overall Performance:**

| Evaluation Metric | AI Model (%) | Rule-Based (%) |
|---|---|---|
| Source Classification | **90** | 78 |
| Bias Detection | **89** | 65 |
| Claim Extraction | **92** | 71 |
| **Overall Accuracy** | **91.7** | 71.3 |

*Table 2: Evaluation TableI*

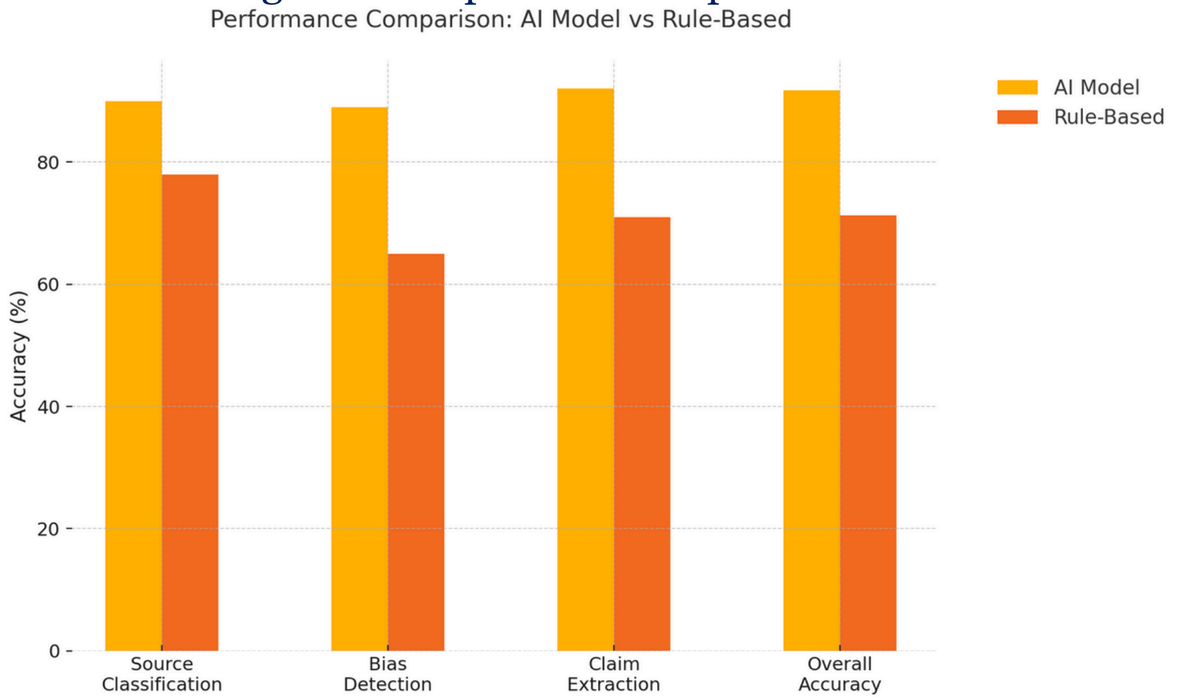
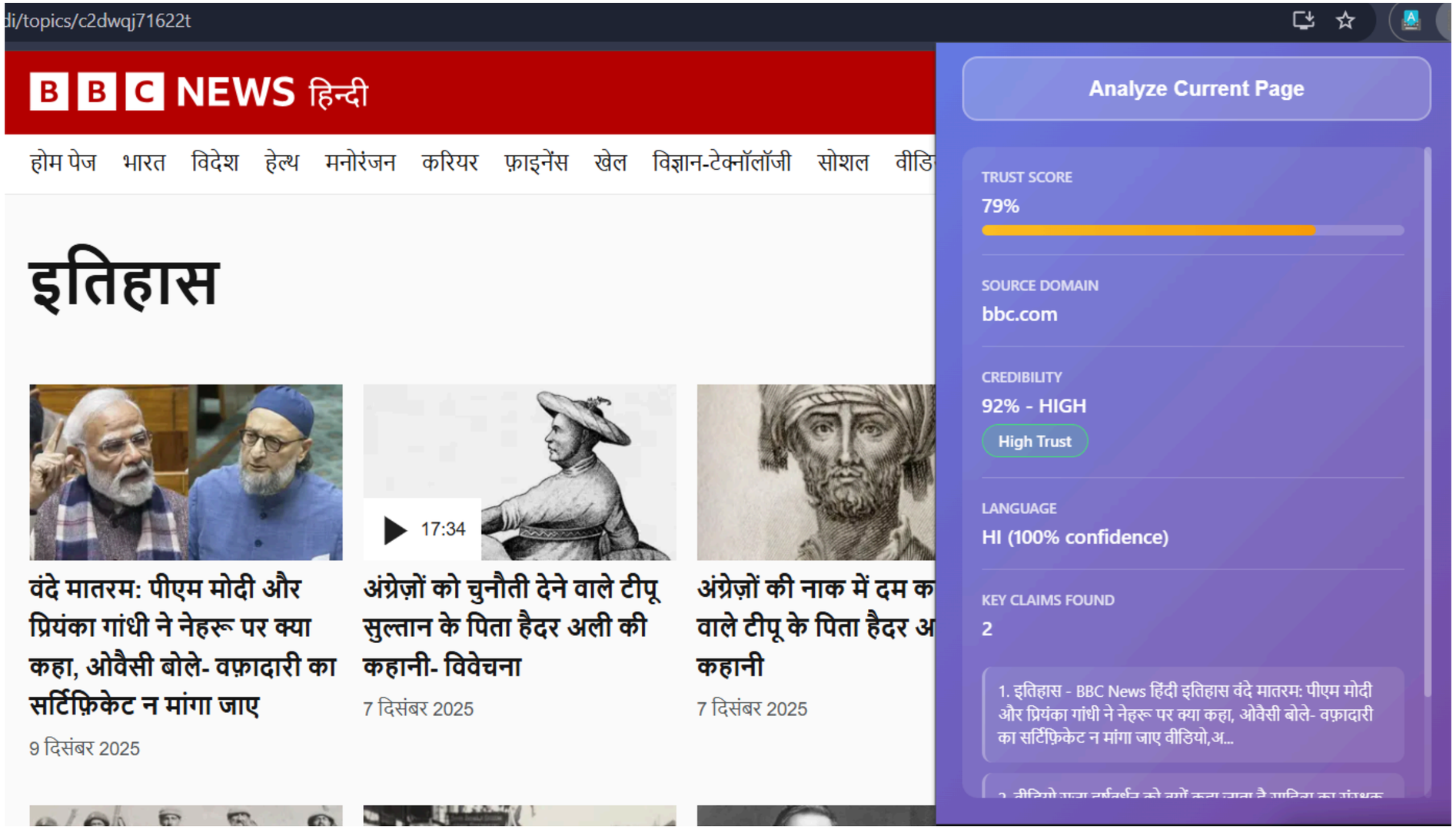
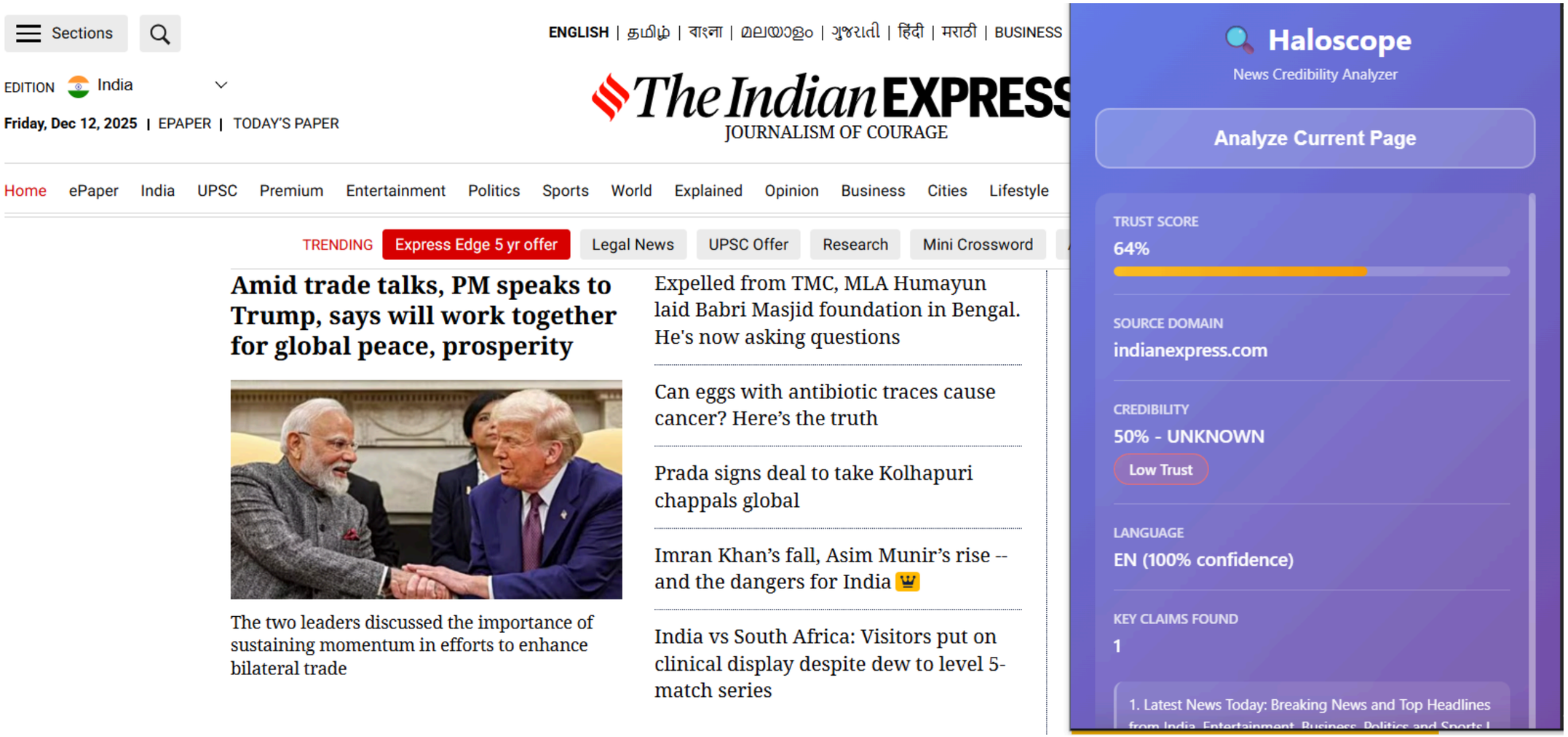



*Fig 4: Evaluation Plott*

# Testing on Real Websites



# Testing on Real Websites



# Conclusion and Future Directions

## WHAT WE ACHIEVED

**91.7%**
Overall Accuracy
(AI Mode)

**12-15s**
Real-Time Analysis
(Production Ready)

### 4-Agent Agentic AI System

ContentAnalyzer | ClaimExtractor | BiasDetector
Orchestrator
Each with AI reasoning + rule-based fallback

## KEY CHALLENGES

### Misleading Content Detection

Half-truths and context manipulation require
deeper semantic understanding beyond surface

### Limited Source Database: Only 13 Sources

Need to expand to 3,500+ sources for coverage

LLM costs require intelligent caching strategies

## RESEARCH IMPACT

**Research Question: Can AI verify as fast as it spreads?**

Speed: YES - 12-15s per article (faster than viral spread)
Accuracy: PARTIAL - 92% on facts, 62.5% on nuance
Explainability: YES - Full reasoning chains provided

### Key Contributions

Multi-agent superiority
validated (+29% accuracy)
Chrome extension
deployed to users

### Technical Innovation

Hybrid AI + rules
(100% uptime)
Explainable verdicts
with audit trails

## FUTURE DIRECTION

### Phase 2

**1. Enhanced Agent Communication**
Agent-to-agent debate protocols | Challenge mechanisms
Add VerificationAgent for web-based fact-checking

**2. Advanced Semantic Analysis**
Integrate BERT/GPT for context understanding
Build knowledge graphs for entity relationships

**3. Scale to Production**
Expand to 5,000+ sources | 80% cache hit rate
Process 10,000 articles/hour at scale

# Thank You


**Dr. Shyama Prasad Mukherjee International Institute of Information Technology, Naya Raipur**

# Literature Review - Research Background

| Study | Technique | Dataset | Accuracy | Key Limitation |
|---|---|---|---|---|
| Shu et al. (2017) | SVM + Decision Trees | BuzzFeed, PolitiFact | 72-78% | Binary only, no nuance detection |
| Wang (2017) LIAR | Hybrid CNN | 12,836 claims | 61% (binary) | Poor on nuanced content |
| Kaliyar (2020) FakeBERT | BERT Deep Learning | ISOT, Kaggle | 98.90% | Overfits, fails on real-world noisy data |
| Cui & Lee (2020) | Multi-modal | CoAID COVID-19 | F1: 91.6% | Domain-specific, doesn't generalize |

Table 1: Comparison of Machine Learning Approaches for Fake News Detection

# Literature Review - Research Background

| Study | Technique | Dataset | Accuracy | Key Limitation |
|---|---|---|---|---|
| Shu (2019) dEFEND | Graph Neural Network | PolitiFact, GossipCop | 89.20% | Needs social network data, expensive |
| Hakak et al. (2021) | Ensemble: NB+RF+SVM | FakeNewsNet, ISOT | 94.20% | Not real-time, degrades on adversarial |
| Monti (2019) | Geometric Deep Learning | BuzzFeed | 88.70% | Doesn't scale, misses content manipulation |
| Zhang & Ghorbani (2020) | Multi-modal: Text+Image | Multiple | 70-92% | Single models fail on sophisticated fakes |

Table 2 : Advanced Machine Learning Techniques for Fake News Detection