PROJECT OSKAR 2.0

A Multimodal Decision-Support Ecosystem for Transparent, Context-Aware Content Moderation

Technical Architecture Document

February 21, 2026

Abstract

PROJECT OSKAR (Online Safety & Knowledge Authenticity Resolver) has been fundamentally reimagined from a modular detection pipeline into a comprehensive *moderation decision-support ecosystem*. This document presents OSKAR 2.0, an architecture emphasizing transparency, human-AI collaboration, and continuous learning. The system integrates multimodal intelligence (text, image, audio), cross-comment contextual analysis, network-level bot detection, and calibrated uncertainty quantification. We introduce novel contributions including dynamic risk scoring with context-dependent weight modulation, conversation graph embeddings for thread-aware classification, and an active learning framework with demographic bias mitigation. OSKAR 2.0 targets deployment as both a real-time API service and a moderator-facing analytics platform, with explicit ethical governance mechanisms and adversarial robustness guarantees.

Contents

1 Introduction **3**
- 1.1 Core Philosophy Shift . . . 3
- 1.2 Key Contributions . . . 3

2 System Architecture **3**
- 2.1 Layer 1: Knowledge & Memory Infrastructure . . . 4
  - 2.1.1 Versioned Knowledge Graph . . . 4
  - 2.1.2 Semantic Caching Layer . . . 4
- 2.2 Layer 2: Multimodal Intelligence Core . . . 4
  - 2.2.1 Multilingual Foundation . . . 4
  - 2.2.2 Conversation Graph Embeddings . . . 5
- 2.3 Layer 3: Cognitive Engine . . . 5
  - 2.3.1 Confidence Calibration . . . 5
  - 2.3.2 Uncertainty Quantification . . . 5
- 2.4 Layer 4: Decision Interface . . . 5
  - 2.4.1 Tiered Intervention Strategy . . . 6

3 Module Specifications **6**
- 3.1 Module A: Socio-Cultural Content Analysis . . . 6
  - 3.1.1 Training Data Stack . . . 6
  - 3.1.2 Model Architecture . . . 6
  - 3.1.3 Output Schema . . . 7
- 3.2 Module B: Structured Claim Understanding . . . 7
  - 3.2.1 Hierarchical Claim Classification . . . 7

3.2.2 Metadata Extraction . . . 7
3.3 Module C: Probabilistic Verification Engine . . . 7
3.3.1 Multi-Hop Evidence Retrieval . . . 8
3.3.2 Source Credibility Weighting . . . 8
3.4 Module D: Network Behavior Analysis . . . 8
3.4.1 Multi-Scale Feature Extraction . . . 8
3.4.2 Coordinated Inauthentic Behavior (CIB) Detection . . . 9
3.5 Module E: Dynamic Risk Fusion . . . 9
3.5.1 Context-Dependent Weight Modulation . . . 9
3.5.2 Uncertainty Propagation . . . 9

4 Continuous Learning Framework **9**
4.1 Active Learning Pipeline . . . 10
4.2 Bias Mitigation Protocol . . . 10

5 Evaluation Framework **10**
5.1 Module-Level Metrics . . . 10
5.2 System-Level Metrics . . . 10
5.3 Adversarial Robustness . . . 11

6 Ethical Governance **11**
6.1 Political Neutrality Policy . . . 11
6.2 Jurisdictional Adaptation . . . 11

7 Implementation Roadmap **12**

8 Conclusion **12**

1 Introduction

The proliferation of misinformation, hate speech, and coordinated inauthentic behavior across digital platforms has created an urgent need for moderation systems that transcend simple classification tasks. Traditional approaches suffer from fundamental limitations: they analyze content in isolation, lack transparency in decision-making, fail to adapt to evolving linguistic patterns, and ignore the multimodal nature of modern harmful content.

PROJECT OSKAR 2.0 addresses these gaps through a paradigm shift from *detection* to *decision support*. Rather than replacing human judgment, OSKAR 2.0 augments moderator capabilities with calibrated uncertainty estimates, explainable evidence chains, and network-level intelligence. The system is designed as a four-layer architecture spanning infrastructure, intelligence, cognition, and interfaces.

1.1 Core Philosophy Shift

Table 1: Evolution from OSKAR 1.0 to OSKAR 2.0

Original OSKAR	**Reimagined OSKAR 2.0**
Pipeline of independent detection modules	Ecosystem of trust, transparency, and human-AI collaboration
Point-in-time content analysis	Temporal, contextual, and network-aware intelligence
English-centric, text-only processing	Multilingual, multimodal, culturally adaptive analysis
Static model deployment	Continuously learning, self-monitoring system
Simple API endpoint	Comprehensive decision-support platform with feedback loops

1.2 Key Contributions

This document makes the following technical contributions:

1. **Multimodal Contextual Intelligence**: Integration of conversation graph embeddings, cross-modal attention mechanisms, and thread-level context windows.
2. **Calibrated Uncertainty Quantification**: Temperature-scaled confidence scores with entropy-based thresholds for explicit "I don't know" states.
3. **Dynamic Risk Fusion**: Context-dependent weight modulation based on claim type, platform state, and user history.
4. **Network-Aware Bot Detection**: Graph neural network analysis of coordinated inauthentic behavior spanning individual, temporal, and network scales.
5. **Ethical Governance Framework**: Built-in bias mitigation, demographic parity auditing, and jurisdictional adaptation mechanisms.

2 System Architecture

OSKAR 2.0 implements a four-layer stack with clear separation of concerns and bidirectional information flow between layers.

2.1 Layer 1: Knowledge & Memory Infrastructure

The foundation layer provides persistent, versioned storage of evidence, embeddings, and decision histories.

2.1.1 Versioned Knowledge Graph

Unlike static Wikipedia snapshots, OSKAR 2.0 maintains temporally-versioned knowledge bases:

```
evidence_corpus/
          wikipedia/
                2024-01-15.snapshot/
                2024-06-15.snapshot/
                latest -> 2024-06-15.snapshot/
          fact_checks/
                politifact_2024.jsonl
                snopes_2024.jsonl
          scientific/
                pubmed_embeddings.faiss
          embeddings/
              wikipedia_2024-01-15.faiss
              wikipedia_2024-06-15.faiss
```

2.1.2 Semantic Caching Layer

Frequent claims are cached using Redis with TTL strategies:

- **Breaking claims**: 1 hour TTL (rapidly evolving)
- **Stable facts**: 7 days TTL (established knowledge)
- **Claim embeddings**: 24 hours TTL (computationally expensive)

2.2 Layer 2: Multimodal Intelligence Core

This layer processes raw content across modalities while maintaining conversational context.

2.2.1 Multilingual Foundation

OSKAR 2.0 replaces DistilBERT with **Gemma-2 2B IT**, selected for:

- Native support for 140+ languages without translation degradation
- 8,192 token context window for thread-level analysis
- Instruction-tuned variant for few-shot claim classification
- CPU-efficient inference suitable for edge deployment

Language detection uses fastText lid.176 with automatic routing:

- High-resource languages: Direct Gemma-2 inference
- Low-resource languages: Adapter-based fine-tuning (LoRA)
- Code-switching: Segmentation and parallel processing

2.2.2 Conversation Graph Embeddings

Comments are represented as nodes in a dynamic graph $\mathcal{G} = (V, E)$ where:

- V represents users, comments, and media elements
- E encodes reply relationships, mentions, and temporal adjacency

Graph Attention Networks (GAT) propagate risk signals:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}\right) \tag{1}$$

where α_{ij} are attention coefficients learned during training.

2.3 Layer 3: Cognitive Engine

The cognitive layer aggregates module outputs into calibrated, explainable decisions.

2.3.1 Confidence Calibration

Raw model probabilities are calibrated using Platt scaling:

$$P(y = 1|x) = \frac{1}{1 + \exp(Af(x) + B)} \tag{2}$$

where $f(x)$ is the model logit, and parameters A, B are fit on a validation set to minimize Expected Calibration Error (ECE).

Temperature scaling refines confidence estimates:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{3}$$

with $T < 1$ for under-confident models, $T > 1$ for over-confident models.

2.3.2 Uncertainty Quantification

OSKAR 2.0 distinguishes *aleatoric* (data) and *epistemic* (model) uncertainty:

Algorithm 1 Uncertainty-Based Routing

Require: Prediction distribution $P(y|x)$, entropy threshold τ

Compute predictive entropy: $H = -\sum_c P(y = c|x) \log P(y = c|x)$
Compute maximum probability: $p_{max} = \max_c P(y = c|x)$
if $H > \tau$ **or** $p_{max} < 0.6$ **then**
    return UNCERTAIN $\rightarrow$ Human review queue
else if $0.4 \leq$ entropy $\leq \tau$ **then**
    return LOWCONFIDENCE $\rightarrow$ Soft warning
else
    return HIGHCONFIDENCE $\rightarrow$ Automated action
end if

2.4 Layer 4: Decision Interface

The interface layer translates model outputs into actionable interventions with full transparency.

2.4.1 Tiered Intervention Strategy

Table 2: Risk-Based Intervention Levels

Risk	Action	System Response
Low (< 0.3)	None	Log only; no user-facing action
Medium ($0.3 - 0.7$)	Pre-post warning	"This claim appears unverified. Here's what we found..." with evidence preview
High ($0.7 - 0.9$)	Flag for review	Inline fact-check displayed; post held for moderator approval
Critical (> 0.9)	Immediate hold	Content blocked; network analysis triggered; urgent moderator alert

3 Module Specifications

3.1 Module A: Socio-Cultural Content Analysis

Objective: Detect hate speech, harassment, and identity-based attacks with cultural context awareness.

3.1.1 Training Data Stack

Table 3: Hate Speech Training Corpus

Dataset	Size	Purpose
HateXplain	20K	Human rationales for explainability
TweetEval (Hate)	9K	Benchmark consistency; irony detection
Davidson (2017)	25K	Baseline comparison
Synthetic Adversarial	50K	Character substitutions, leetspeak, diacritics
AAVE Samples	10K	Bias mitigation; dialectal variation
Cultural Context	15K	Reclaimed vs. targeted slur disambiguation

3.1.2 Model Architecture

Primary model: **Gemma-2 2B IT** with classification head.

Auxiliary sarcasm detector: RoBERTa-large fine-tuned on SARC dataset, processing:

- Text sentiment (VADER scores)
- Emoji patterns and density
- Punctuation anomalies (!!!, ???, mixed case)
- Contrastive sentiment (text vs. image caption)

3.1.3 Output Schema

```
{
  "hate_label": true,
  "severity_score": 0.87,
  "target_categories": ["race", "religion"],
  "rationale": "Attention weights on 'targeted slur'",
  "cultural_context_flag": false,
  "sarcasm_detected": true,
  "confidence": 0.91,
  "calibration": "well-calibrated"
}
```

3.2 Module B: Structured Claim Understanding

Objective: Identify verifiable claims and route to appropriate evidence sources.

3.2.1 Hierarchical Claim Classification

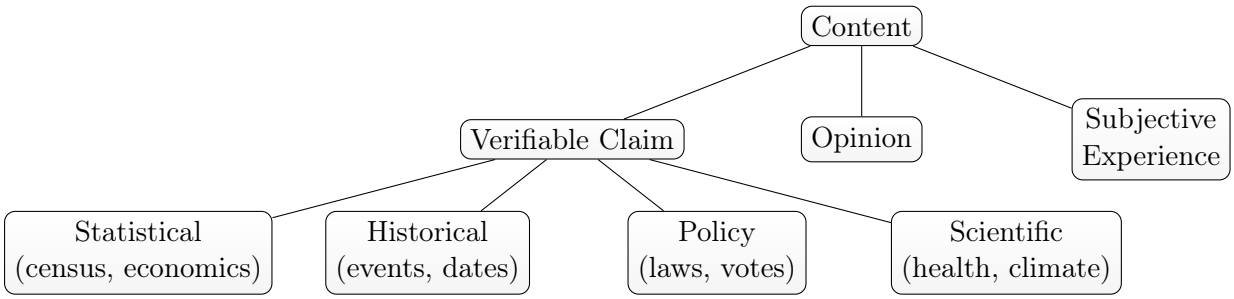

Figure 1: Claim Type Taxonomy

3.2.2 Metadata Extraction

Named Entity Recognition (NER) extracts:

- **WHO**: Persons, organizations (link to knowledge graph)
- **WHAT**: Key events, policies, statistics
- **WHEN**: Temporal scope for evidence retrieval
- **WHERE**: Geopolitical relevance for localized fact-checking

3.3 Module C: Probabilistic Verification Engine

Objective: Retrieve evidence and classify claim veracity with calibrated uncertainty.

3.3.1 Multi-Hop Evidence Retrieval

Algorithm 2 Iterative Evidence Retrieval

Require: Claim c, Knowledge bases $\mathcal{K} = \{K_1, ..., K_n\}$, Max hops H
Initialize evidence set $\mathcal{E} = \emptyset$
Embed claim: $\mathbf{e}_c = \text{SBERT}(c)$
for $h = 1$ to H **do**
  for each $K_i \in \mathcal{K}$ **do**
    Retrieve top-k: $\mathcal{R}_i = \text{FAISS}(\mathbf{e}_c, K_i, k)$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{R}_i$
  end for
  Extract entities from $\mathcal{E}$: $\mathcal{N} = \text{NER}(\mathcal{E})$
  Query knowledge graph: $\mathcal{G}_n = \text{Neo4j}(\mathcal{N})$
  Expand claim with relations: $c \leftarrow c \oplus \mathcal{G}_n$
end for
return $\mathcal{E}$ ranked by relevance and source credibility

3.3.2 Source Credibility Weighting

Evidence sources are weighted by Media Bias/Fact Check ratings:

$$w_{source} = \begin{cases} 1.0 & \text{Very High} \\ 0.8 & \text{High} \\ 0.6 & \text{Mostly Factual} \\ 0.3 & \text{Mixed} \\ 0.1 & \text{Low} \\ 0.0 & \text{Very Low / Conspiracy} \end{cases} \tag{4}$$

3.4 Module D: Network Behavior Analysis

Objective: Detect automated accounts and coordinated inauthentic behavior across multiple scales.

3.4.1 Multi-Scale Feature Extraction

Table 4: Bot Detection Feature Hierarchy

Scale	Features	Model	Target
Individual	Posting velocity, content similarity, linguistic diversity (MTLD)	XGBoost	Automated posting
Temporal	Burst patterns, circadian entropy, inter-post timing	LSTM autoencoder	Scheduled automation
Network	Coordination clusters, shared embedding spaces, synchronized actions	GraphSAGE	Coordinated campaigns
Content	Perplexity scores, stylometry, semantic coherence	Fine-tuned RoBERTa	LLM-generated text

3.4.2 Coordinated Inauthentic Behavior (CIB) Detection

1. Embed all posts in semantic space: $\mathbf{E} = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$ where $\mathbf{e}_i = \text{SBERT}(\text{post}_i)$
2. Online clustering using HDBSCAN with dynamic ϵ
3. Flag clusters meeting criteria:
   - Minimum 5 accounts
   - Temporal proximity ($\Delta t < 1$ hour)
   - Low account diversity (new accounts, similar bios)
   - Narrative consistency (cosine similarity > 0.85)

3.5 Module E: Dynamic Risk Fusion

Objective: Aggregate module outputs into context-sensitive risk scores with uncertainty propagation.

3.5.1 Context-Dependent Weight Modulation

Algorithm 3 Dynamic Risk Calculation

Require: Module scores $\mathbf{s} = [s_{mis}, s_{hate}, s_{bot}]$, Context $\mathcal{C}$, User history $\mathcal{H}$
Initialize base weights: $\mathbf{w} = [0.4, 0.3, 0.3]$
if claim_type = POLITICAL $\wedge$ platform_context = ELECTIONSEASON **then**
    $w_{mis} \leftarrow w_{mis} \times 1.5$
    $w_{hate} \leftarrow w_{hate} \times 1.2$
end if
if user_history.previous_flags > 2 **then**
    $w_{hate} \leftarrow w_{hate} \times 1.3$
end if
if thread_depth $> 5 \wedge$ toxicity_escalation **then**
    $w_{hate} \leftarrow w_{hate} \times 1.4$
end if
Normalize: $\mathbf{w} \leftarrow \mathbf{w} / \sum w_i$
return MonteCarloSimulation($\mathbf{s}, \mathbf{w}, n = 1000$)

3.5.2 Uncertainty Propagation

Risk distributions are estimated via Monte Carlo simulation:

$$\mathcal{R} = \{r^{(i)} = \mathbf{w} \cdot \tilde{\mathbf{s}}^{(i)}\}_{i=1}^{N} \tag{5}$$

where $\tilde{\mathbf{s}}^{(i)}$ are samples from module uncertainty distributions.

Output includes:

- Mean risk score: $\bar{r} = \frac{1}{N} \sum r^{(i)}$
- 95% confidence interval: $[\hat{r}_{0.025}, \hat{r}_{0.975}]$
- Calibration status: comparison of predicted vs. observed accuracy

4 Continuous Learning Framework

OSKAR 2.0 implements a closed-loop learning system ensuring model freshness and bias mitigation.

4.1 Active Learning Pipeline

1. **Uncertainty Sampling**: Select predictions with entropy $H > 0.8$
2. **Diversity Sampling**: Ensure coverage via k-Means++ on embedding space
3. **Human Review**: Expert moderators label selected cases with rationale
4. **Model Update**: Weekly LoRA fine-tuning ($r = 16, \alpha = 32$)
5. **Evaluation**: Rolling benchmark on last 30 days; rollback if F1 drops $> 2\%$

4.2 Bias Mitigation Protocol

Table 5: Demographic Parity Monitoring

Stage	Action	Threshold
Pre-deployment	Evaluate on AAVE, Indian English, Hispanic English test sets	Disparity $< 10\%$ in F1
Ongoing monitoring	Track flag rates by demographic proxies (language patterns, geolocation)	Monthly audit reports
Correction trigger	Rebalance training data if disparity exceeds threshold	Stratified sampling
Transparency	Public bias audit reports	Quarterly publication

5 Evaluation Framework

5.1 Module-Level Metrics

Table 6: Evaluation Metrics by Module

Module	Primary	Secondary	Fairness
Hate Speech	F1-macro	AUC-ROC	Demographic parity across dialects
Claim Detection	Accuracy	Per-class F1	Language group parity
Verification	ECE (Calibration)	Precision@K	Source diversity index
Bot Detection	ROC-AUC	AUPRC	False positive by account age

5.2 System-Level Metrics

- **Harm Reduction**: Percentage of misinformation corrected before viral threshold (100 shares)
- **False Positive Harm**: Estimated user impact score incorporating appeal success rate
- **Latency**: P50, P95, P99 response times by modality
- **Coverage**: Percentage of content analyzed vs. sampled
- **Human-AI Agreement**: Cohen's κ between model and expert moderator consensus

5.3 Adversarial Robustness

Table 7: Adversarial Attack Defenses

Attack	Method	Defense
Character substitution	"v4cc1n3s", "h4t3"	Fuzzy matching + CharCNN embeddings
Invisible characters	Zero-width joiners, homoglyphs	Unicode normalization + sanitization
Synonym replacement	WordNet adversaries	Robust paraphrase detection (SBERT)
Image-based claims	Memes with false text overlay	OCR (Tesseract) + visual-textual consistency check
Coordinated campaigns	Synthetic bot networks	GNN clustering + temporal burst detection
Dialectal variation	AAVE, regional dialects	Dialect-robust pre-training

6 Ethical Governance

6.1 Political Neutrality Policy

OSKAR 2.0 adheres to strict neutrality principles:

- Verifies only *empirically verifiable claims* using dataset-driven evidence
- Does not judge *opinions*, *values*, or *policy preferences*
- Provides *calibrated confidence scores* with explicit uncertainty
- Ensures all interventions are *explainable* and *appealable*

6.2 Jurisdictional Adaptation

- **EU (DSA compliance)**: Strict thresholds, mandatory human review for political content
- **US (First Amendment)**: High thresholds for hate speech, emphasis on counter-speech
- **Global South**: Localized fact-check sources, low-resource language support

7 Implementation Roadmap

Table 8: Development Phases

Phase	Timeline	Deliverables
1. Foundation	Months 1–2	Gemma-2 2B integration, confidence calibration, Redis caching, SHAP explainability, Tesseract OCR
2. Intelligence	Months 3–4	Thread context analysis, claim type classification, uncertainty thresholds, active learning pipeline, moderator dashboard v1
3. Scale	Months 5–6	Multilingual adapters (Hindi, Spanish, Arabic), GNN bot detection, knowledge graph integration, browser extension
4. Ecosystem	Months 7–9	Narrative clustering, polarization index, cross-platform APIs, enterprise on-prem deployment, full audit framework

8 Conclusion

PROJECT OSKAR 2.0 represents a fundamental evolution in content moderation technology. By integrating multimodal intelligence, network-aware detection, calibrated uncertainty, and ethical governance into a unified ecosystem, OSKAR 2.0 addresses the critical limitations of existing approaches. The architecture emphasizes transparency, human oversight, and continuous improvement, positioning it as both a research contribution and a practical solution for platform governance.

Future work will explore federated learning for privacy-preserving multi-platform training, real-time multilingual speech analysis, and integration with decentralized identity systems for enhanced bot detection.

OSKAR Online Safety & Knowledge Authenticity Resolver

GAT Graph Attention Network

GNN Graph Neural Network

LoRA Low-Rank Adaptation

ECE Expected Calibration Error

CIB Coordinated Inauthentic Behavior

AAVE African American Vernacular English

NER Named Entity Recognition

TTL Time To Live