THE REIMAGINED OSKAR: A Moderation Decision-Support Ecosystem

Core Philosophy Shift

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

I. ARCHITECTURE 2.0: The Four-Layer Stack

```
┌──────────────────────────────────────────────────────────────
────────┐
│  LAYER 4: DECISION INTERFACE                          │
│  • Pre-post warning system                            │
│  • Moderator dashboard with heatmaps & audit trails    │
│  • Explainability panel (attention maps, evidence chains)   │
│  • Appeal mechanism with human override              │
└──────────────────────────────────────────────────────────────
────────┘
                        ↕
┌──────────────────────────────────────────────────────────────
────────┐
│  LAYER 3: COGNITIVE ENGINE                            │
│  • Confidence Calibration Layer (Platt scaling, temperature) │
│  • Uncertainty Quantification (entropy-based thresholds)     │
│  • Dynamic Risk Scoring (context-dependent weights)          │
│  • Claim Type Classification (statistical/historical/policy) │
└──────────────────────────────────────────────────────────────
────────┘
                        ↕
┌──────────────────────────────────────────────────────────────
────────┐
│  LAYER 2: MULTIModal INTELLIGENCE CORE                 │
│  • Cross-Comment Context (thread embeddings)            │
│  • Multilingual Foundation (Gemma-2 2B + adapters)      │
│  • Multimodal Fusion (text + image OCR + audio transcription) │
│  • Narrative Clustering (coordinated inauthentic behavior)   │
└──────────────────────────────────────────────────────────────
────────┘
                        ↕
```

┌──
──────┐
│ LAYER 1: KNOWLEDGE & MEMORY INFRASTRUCTURE │
│ • Versioned Knowledge Graph (Wikipedia + fact-checks + news) │
│ • Semantic Cache (Redis for claim embeddings) │
│ • Feedback Learning Database (human corrections → model updates)│
│ • Audit-Grade Logging (immutable decision records) │
└──
──────┘
```

---

## **II. MODULE-BY-MODULE TRANSFORMATION**

### **A. Preprocessing Layer → Contextual Intelligence Hub**

**Original**: Basic text cleaning
**Upgrade**:

| Feature | Implementation | Impact |
|---------|---------------|--------|
| **Language Detection** | fastText lid.176 → route to appropriate model | Handles code-switching, regional dialects |
| **Thread Reconstruction** | Graph traversal of parent/child comments | Context-aware classification |
| **Temporal Metadata** | Posting time, edit history, velocity | Detects burst behavior |
| **Media Extraction** | OCR (Tesseract) + Audio (Whisper) + Keyframes | Multimodal grounding |

**Key Innovation**: **Conversation Graph Embeddings**
- Represent threads as graphs (users → comments → reactions)
- Use Graph Attention Networks (GAT) to propagate risk signals through the network
- A suspicious reply to a clean post inherits elevated scrutiny

---

### **B. Hate Speech Detection → Socio-Cultural Content Analysis**

**Original**: DistilBERT on Davidson dataset
**Upgrade**: **Gemma-2 2B IT** with:

```
Training Data Stack:
├── HateXplain (human rationales for explainability)
├── TweetEval (irony, sentiment, hate benchmarks)
├── AAVE-inclusive samples (bias mitigation)
├── Synthetic adversarial examples (character substitutions, leetspeak)

└── Cultural context annotations (reclaimed vs. targeted slurs)

Output:
├── Hate label (binary)
├── Severity score (0-1)
├── Target identity categories (protected classes)
├── Rationale extraction (attention-based)
└── Cultural context flag (requires human review)
```

Critical Addition: **Sarcasm & Irony Detection**
- Use emoji patterns, punctuation anomalies, and contrastive sentiment (text vs. image)
- Separate pipeline: RoBERTa-large trained on SARC dataset

C. Claim Detection → Structured Claim Understanding

Original: Binary claim/non-claim
Upgrade: **Hierarchical Claim Typing**

```
Claim Classification Tree:
├── Verifiable Claim
│   ├── Statistical Claim → Query: data.gov, World Bank, scientific DBs
│   ├── Historical Claim  → Query: Wikipedia (versioned), archives
│   ├── Policy Claim      → Query: Official press releases, voting records
│   └── Scientific Claim  → Query: PubMed, Semantic Scholar
├── Opinion (skip verification)
└── Subjective Experience (skip verification)

Metadata Extracted:
├── Entities (NER: persons, organizations, locations)
├── Temporal scope (when did this happen?)
├── Geopolitical relevance (affects scoring weights)
└── Source cited in original post (for source credibility check)
```

D. Evidence Retrieval → Knowledge Graph + Live Search Hybrid

Original: SBERT + FAISS + Wikipedia
Upgrade: **Multi-Hop Evidence Retrieval**

Component	Technology	Purpose
Vector Store	FAISS + ScaNN (Google)	Billion-scale similarity search
```

Knowledge Graph	Neo4j	Entity-relationship verification (e.g., "X is CEO of Y")
Live Search	SerpAPI + Fact-check APIs	Breaking claims, recent events
Source Credibility	MediaBias/FactCheck integration	Weight evidence by source reliability
Temporal Versioning	Wikipedia snapshots + news timestamps	Handle evolving truths

Retrieval Strategy by Claim Type:
- **Statistical**: Prioritize primary sources (census, official statistics)
- **Historical**: Prioritize academic consensus, multiple corroborating sources
- **Policy**: Prioritize official government documents, voting records
- **Scientific**: Prioritize peer-reviewed, recent meta-analyses

---

E. Verification Engine → Probabilistic Inference System

Original: NLI (RoBERTa-large-MNLI) → Supported/Refuted/NEI
Upgrade: **Uncertainty-Aware Verification**

```
Verification Output:
├── Verdict Distribution:
│   ├── Supported: 0.65
│   ├── Refuted: 0.25
│   ├── Not Enough Info: 0.10
│   └── Confidence: 0.40 (entropy-based uncertainty)
├── Evidence Chain: [Source A] → [Source B] → [Conclusion]
├── Confidence Calibration: Temperature-scaled probabilities
└── Knowledge Cutoff: "Evidence current as of 2024-06-15"
```

Decision Logic:
- If max_probability < 0.6 → **"Uncertain"** (flag for human review)
- If evidence_conflict > threshold → **"Disputed"** (present multiple viewpoints)

F. Bot Detection → Network Behavior Analysis

Original: Isolation Forest on basic features
Upgrade: **Multi-Scale Bot Detection**

Scale	Features	Model
Individual	Posting velocity, content similarity, linguistic diversity	XGBoost
Temporal	Burst patterns, circadian rhythm analysis, inter-post timing entropy	LSTM autoencoder

| **Network** | Coordination clusters, shared embedding spaces, synchronized posting | Graph Neural Network (GNN) |
| **Content** | LLM-generated text detection (perplexity + stylometry) | Fine-tuned detector |

Coordinated Inauthentic Behavior (CIB) Detection:
- Embed all posts in semantic space (SBERT)
- Online clustering (HDBSCAN) to find narrative clusters
- Flag: Same claim, multiple accounts, temporal proximity, low account diversity

---

G. Risk Scoring → Dynamic, Explainable Risk Engine

Original: Static weighted sum
Upgrade: **Context-Adaptive Risk Fusion**

```python
def calculate_risk(claim_type, platform_context, user_history):
    base_weights = {
        'misinformation': 0.4,
        'hate_speech': 0.3,
        'bot_likelihood': 0.3
    }

    # Context-dependent modulation
    if claim_type == 'political_claim' and platform_context == 'election_season':
        base_weights['misinformation'] *= 1.5
        base_weights['hate_speech'] *= 1.2

    if user_history['previous_flags'] > 2:
        base_weights['hate_speech'] *= 1.3  # Escalation pattern

    # Normalize
    total = sum(base_weights.values())
    weights = {k: v/total for k, v in base_weights.items()}

    # Monte Carlo simulation for uncertainty propagation
    risk_distribution = monte_carlo_sim(scores, weights, n=1000)

    return {
        'risk_level': 'High',
        'mean_score': 0.85,
        'confidence_interval': [0.78, 0.91],
        'calibration_status': 'well-calibrated',
        'contributing_factors': [
            {'factor': 'misinformation', 'contribution': 0.45, 'evidence': '...'},
            {'factor': 'bot_likelihood', 'contribution': 0.30, 'evidence': '...'}
        ]
```

}
```

---

### **H. Response Generator → Interactive Correction System**

**Original**: Template-based text
**Upgrade**: **Tiered Intervention Strategy**

| Risk Level | User Action | System Response |
|------------|-------------|-----------------|
| **Low** | None | Log only |
| **Medium** | Pre-post warning | "This claim appears unverified. Here's what we found..." |
| **High** | Flag for review + Show correction | Inline fact-check with evidence |
| **Critical** | Hold for moderation + Notify mods | Immediate escalation, network analysis |

**Explainability Panel** (for moderators):
```
┌─────────────────────────────────────┐
│ OSKAR Analysis Report   │
├─────────────────────────────────────┤
│ Claim: "Vaccines cause autism"   │
│ Type: Scientific (Statistical)   │
│   │
│ Risk Score: 0.92 [High]   │
│ Confidence: 0.89 (Well-calibrated) │
│   │
│ Breakdown: │
│ ├── Misinformation: 0.95   │
│ │   └── Evidence: Contradicts CDC, WHO │
│ ├── Hate Speech: 0.05   │
│ └── Bot Probability: 0.15   │
│   │
│ Evidence Chain: │
│ [1] CDC Study (2023) → Refutes   │
│ [2] Lancet Retraction (2010) → Context │
│   │
│ Model Attention: "autism" + "cause" │
│   │
│ [Override] [Request Human Review]   │
└─────────────────────────────────────┘
```

---

## **III. DATA INFRASTRUCTURE & CONTINUOUS LEARNING**
```

Knowledge Base Versioning
```
evidence_corpus/
├── wikipedia/
│   ├── 2024-01-15.snapshot/
│   ├── 2024-06-15.snapshot/
│   └── latest -> 2024-06-15.snapshot/
├── fact_checks/
│   ├── politifact_2024.jsonl
│   └── snopes_2024.jsonl
└── embeddings/
    ├── wikipedia_2024-01-15.faiss
    └── wikipedia_2024-06-15.faiss
```

Active Learning Pipeline
1. **Uncertainty Sampling**: Select predictions with entropy > 0.8
2. **Diversity Sampling**: Ensure coverage of underrepresented claim types
3. **Human Review**: Expert moderators label selected cases
4. **Model Update**: Weekly LoRA fine-tuning on new data
5. **Evaluation**: Rolling benchmark on last 30 days (prevent catastrophic forgetting)

IV. EVALUATION FRAMEWORK: Beyond Accuracy

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

System-Level Metrics
- **Harm Reduction**: % of misinformation corrected before viral spread
- **False Positive Harm**: Estimated user impact of incorrect flags (appeal success rate)
- **Latency**: P50, P95, P99 response times
- **Coverage**: % of content analyzed (vs. sampled)
- **Human-AI Agreement**: Cohen's kappa between model and expert moderators

Adversarial Robustness Testing
Attack	Method	Defense Verified
Character substitution	"v4cc1n3s"	Fuzzy matching + char-level embeddings
Invisible characters	Zero-width joiners	Input sanitization
Synonym replacement	WordNet adversaries	Robust paraphrase detection
Image-based claims	Meme with false text	OCR + visual-textual consistency

| Coordinated campaigns | Synthetic bot networks | GNN clustering detection |

V. PRODUCT FEATURES: From API to Platform

Pre-Post Warning System (Browser Extension)
- Real-time analysis as user types
- Soft warning: "This claim is disputed by [source]"
- Hard warning: "This contains verified misinformation" (requires extra click to post)

Moderator Command Center
- **Heatmap**: Misinformation trends by topic, geography, time
- **Network Graph**: Visualize coordinated bot clusters
- **Audit Trail**: Every decision, override, and appeal logged immutably
- **A/B Testing**: Test different warning messages for effectiveness

Appeal & Feedback System
```
User Flow:
1. Receive notification: "Your post was flagged"
2. View explanation panel with evidence
3. Choose: [Accept] [Appeal] [Edit & Resubmit]
4. If appeal: Routed to human moderator queue
5. If overturned: Model updated via active learning
```

VI. ETHICAL GOVERNANCE FRAMEWORK

Political Neutrality Policy
- OSKAR verifies **verifiable claims** using **dataset-driven evidence**
- Does not judge **opinions**, **values**, or **policy preferences**
- Confidence scores are **transparent** and **calibrated**
- All interventions are **explainable** and **appealable**

Bias Mitigation Protocol
1. **Pre-deployment**: Test on diverse dialects (AAVE, Indian English, etc.)
2. **Ongoing**: Monitor flag rates by demographic proxies
3. **Correction**: Rebalance training data if disparity > 10%
4. **Transparency**: Public bias audit reports quarterly

Jurisdictional Adaptation
- Configurable thresholds by region (EU: strict, US: permissive)
- Localized fact-check sources
- Legal compliance markers (GDPR, DSA)

VII. TECHNOLOGY STACK (Revised)

Layer	Original	Upgraded
Core LLM	DistilBERT	Gemma-2 2B IT (8K context, multilingual)
Embeddings	SBERT	E5-large + domain-adapted versions
Vector DB	FAISS	FAISS + ScaNN + Redis caching
Graph DB	None	Neo4j (knowledge graph)
Orchestration	Sequential	Asyncio + Celery + Ray (distributed)
Monitoring	None	Prometheus + Grafana + MLflow
Deployment	Docker	Kubernetes + auto-scaling

VIII. IMPLEMENTATION ROADMAP

Phase 1: Foundation (Months 1-2)
- [ ] Replace DistilBERT with Gemma-2 2B
- [ ] Implement confidence calibration (Platt scaling)
- [ ] Add Redis caching layer
- [ ] Build basic explainability panel (SHAP)
- [ ] Integrate Tesseract OCR

Phase 2: Intelligence (Months 3-4)
- [ ] Thread context analysis (conversation graphs)
- [ ] Claim type classification
- [ ] Uncertainty threshold system
- [ ] Active learning pipeline
- [ ] Moderator dashboard v1

Phase 3: Scale (Months 5-6)
- [ ] Multilingual adapters (Hindi, Spanish, Arabic)
- [ ] Advanced bot detection (GNN)
- [ ] Knowledge graph integration
- [ ] Pre-post warning browser extension
- [ ] Adversarial robustness testing suite

Phase 4: Ecosystem (Months 7-9)
- [ ] Real-time narrative clustering
- [ ] Polarization index calculation
- [ ] Cross-platform integration APIs
- [ ] Enterprise deployment (on-prem option)
- [ ] Full audit & compliance framework

IX. DIFFERENTIATION: Why OSKAR 2.0 Wins

Competitor Approach	OSKAR 2.0 Advantage
OpenAI Moderation API	Transparent, auditable, on-premise capable
Google Perspective	Multimodal, thread-aware, explainable
Hive Moderation	Academic rigor, continuous learning, bias mitigation
In-house platform systems	Modular, API-first, human-in-the-loop design

---

X. SUCCESS METRICS (12-Month Targets)

- **Accuracy**: F1 > 0.90 on hate speech, >0.85 on claim verification
- **Calibration**: Expected Calibration Error < 0.05
- **Latency**: P95 < 200ms for text, < 2s for multimodal
- **Coverage**: 50+ languages supported
- **Adoption**: 3 pilot platforms, 10M+ analyzed posts
- **Trust**: 80% moderator agreement, <5% appeal rate
