Multi Agent Doctrine Synthesis for Autonomous Geopolitical Forecasting, The MASX AI Methodology

Research Paper, Technical Preprint April 2026

.

Abstract

This paper presents MASX AI, a **fully autonomous** strategic forecasting system that generates probabilistic geopolitical predictions through a novel **multi agent doctrine synthesis** architecture, operating from raw media ingestion through multi channel distribution without human intervention. The system is foundation is a **14 stage data pipeline** that ingests thousands of articles daily from 200+ sources, applies 4 gate progressive corroboration (with syndication detection via MinHash, bias risk guards, and multi scale substance validation), cross day lifecycle tracking via centroid embedding matching, and a dual provider LLM significance judge. Corroborated events are then analyzed through a corpus of 35 strategic doctrines, spanning classical statecraft (Sun Tzu, Kautilya) through modern hybrid warfare and economic theory, with contextual passages retrieved via per doctrine vector indexes, and probability estimates are synthesized using a suite of deterministic mathematical tools including Bayesian updating, Weibull based temporal projection, Goldstein scaled escalation indexing, and Fermi decomposition. An 8 point grounding validator with auto heal capabilities ensures output quality. The system is most strategically significant innovation is its **self calibrating feedback loop**. Resolved forecasts are scored via Brier decomposition, and the resulting per doctrine performance weights feed back through three independent channels (mathematical weight adjustment, domain specific prompt guidance, and error learning from missed signals), creating a compounding accuracy advantage that deepens with every daily pipeline cycle. The advisor agent produces **quantified intervention windows** and **cost of inaction risk multipliers**, unique decision support outputs not found in any competing platform. This paper details the complete methodology, mathematical foundations, and quality assurance framework.

.

1. Introduction

1.1 The Challenge of Geopolitical Forecasting

Geopolitical forecasting is among the most challenging prediction tasks. Low base rates, complex causal chains, adversarial dynamics, and deep uncertainty all conspire against accuracy. Philip Tetlock is research with the Good Judgment Project demonstrated that disciplined forecasters who update frequently, decompose questions, and consider multiple perspectives significantly outperform both chance and domain experts (Tetlock and Gardner, 2015).

MASX AI operationalizes these principles in software. The system does not rely on a single analytical framework. It selects multiple strategic lenses dynamically based on event characteristics, retrieves relevant doctrine passages through semantic search, constrains LLM synthesis with deterministic mathematical tools, validates output quality through automated grounding checks, and self corrects through calibration feedback from resolved predictions. The entire chain runs autonomously.

1.2 Contributions

This paper makes the following contributions.

A **14 stage autonomous data pipeline** with 4 gate progressive corroboration (syndication detection via MinHash, bias risk guards, multi scale substance validation), 3 tier cross domain detection (keyword to embedding to zero shot), and cross day lifecycle tracking via centroid embedding matching.

A **fully autonomous forecasting pipeline** that operates from signal ingestion through multi channel distribution, resolution verification, and self calibration without human intervention.

A **multi doctrine RAG synthesis** architecture that combines insights from 35 distinct strategic traditions into weighted probabilistic estimates.

A **tool augmented forecasting agent** that constrains LLM output with Bayesian mathematics, temporal survival analysis, and evidence weighting.

An **8 point grounding validator** with deterministic auto heal capabilities for common LLM failure modes.

A **self calibrating feedback loop** that computes per doctrine Brier decomposition and feeds performance weights back through three independent channels, creating a compounding competitive advantage.

Quantified decision support via intervention windows (you have roughly X days) and cost of inaction risk multipliers (waiting costs Y times more risk).

A **dual quality gate architecture** combining 8 point forecast validation with 8 point recommendation validation (16 total checks), both with deterministic auto heal.

A **self improving question quality system** where a 36 term vague phrase blocklist evolves through three tiers based on empirical resolution failure data.

Autonomous multi channel distribution including deterministic newsletter (7 data charts), LinkedIn posts, flagship analyses, and fully autonomous podcast generation (script to TTS to MP3).

Coverage confidence and anomaly detection across 17 geographic regions with censorship, blackout, and surge detection.

.

2. Related Work

2.1 Superforecasting and Calibration

Tetlock is superforecasting research established that calibration, the alignment between predicted probabilities and actual frequencies, is the hallmark of good probabilistic judgment. The Brier score, defined as

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2$$

where p_i is the predicted probability and $o_i \in \{0,1\}$ is the outcome, provides a proper scoring rule where lower scores indicate better calibration. MASX AI uses Brier decomposition (reliability, resolution, uncertainty) to diagnose systematic forecasting errors.

2.2 Multi Agent LLM Systems

Recent work on multi agent LLM architectures demonstrates that deliberation among specialized agents can improve reasoning quality. MASX AI extends this approach by grounding each agent is perspective in a specific strategic corpus (rather than generic prompting), and by mathematically aggregating their outputs through log odds synthesis rather than simple averaging.

2.3 Retrieval Augmented Generation

RAG systems improve LLM factuality by grounding generation in retrieved documents. MASX AI employs 35 independent vector indexes (one per doctrine), each constructed from source PDFs through OCR, semantic chunking, and metadata enrichment. This per doctrine isolation ensures that retrieval quality is not diluted by cross corpus interference.

2.4 AI Forecasting Benchmarks

The Metaculus FutureEval platform and ForecastBench benchmark have established competitive evaluation frameworks for AI forecasting bots, demonstrating that LLM based forecasters are closing in on superforecaster accuracy (Brier gap roughly 0.017 as of early 2026). These systems typically forecast individual questions using web search and ensemble methods. MASX AI differs fundamentally in scope. Rather than competing on individual question accuracy within a tournament framework, it operates the complete chain from signal ingestion through doctrine grounded analysis, question generation, forecasting, quality validation, multi channel distribution, resolution verification, and self calibration, autonomously and daily.

.

3. Methodology

3.1 Data Acquisition, The 14 Stage Data Pipeline

MASX AI operates on **hotspots**, geopolitical events identified and scored by the data pipeline, a **14 stage, 4 gate autonomous data pipeline** that ingests thousands of articles daily from 200+ sources across 17 geographic regions. This pipeline is a critical research contribution in its own right. No competing system implements comparable multi gate corroboration with syndication detection and lifecycle tracking.

3.1.1 Ingestion Architecture (s01 through s03)

Feed Ingestion. RSS feeds are fetched through a 4 layer fallback chain.

(1) Direct RSS with conditional GET (ETag and If Modified Since),

(2) Google News proxy with region aware edition selection,

(3) RSS Bridge,

(4) GDELT DOC backfill.

A circuit breaker skips sources after 5 consecutive failures. Structured APIs (ACLED, GDELT, USGS, NASA EONET, NASA FIRMS, Cloudflare Radar, NGA MSI, GDACS, OONI)

provide supplementary structured data with graceful degradation when API keys are unavailable. Additional sources (FRED, EIA, OTX, URLhaus) are architecturally planned with URL constants and registry entries prepared for future integration.

Fetch Escalation. Article HTML is downloaded through a 7 tier escalation chain (httpx to curl_cffi TLS impersonation to proxy rotation to headless browser to Google Cache) with a 120 second hard budget per URL. SSRF defense in depth applies pre flight DNS resolution with IP validation against private, reserved, and loopback ranges, including IPv4 mapped IPv6 unwrapping. An auto learning domain mode tracker promotes domains through tiers after consecutive failures.

3.1.2 NLP Enrichment and Cross Domain Detection (s04)

Named entity recognition runs via ONNX Runtime on CPU (no GPU required), using a distilbert NER model with custom BIO sequence decoding. A **3 tier cross domain detection** system tags articles with secondary domain classifications.

Tier	Method	Latency	Approach
1	Static keyword rules	roughly 0 ms	Bidirectional matching from cross_domain_rules table
2	Embedding cosine	roughly 2 ms	Article embedding vs 122 domain node embeddings
3	Zero shot classifier	roughly 50 ms	DeBERTa v3 xsmall confirmation for borderline Tier 2

All three tiers run on CPU at $0 inference cost.

3.1.3 Three Layer Deduplication (s05)

Deduplication operates at three levels. (1) Cross run URL hash against all known articles, (2) In run SHA 256 content hash (title + description), (3) SimHash near duplicate detection with 64 bit fingerprints and Hamming distance threshold of 3, catching paraphrased syndication copies.

3.1.4 Four Gate Corroboration Architecture (s06 through s11)

The pipeline is core quality mechanism applies progressive validation through four deterministic gates.

Gate 1 (s06). Junk filter, minimum content length, excluded topic keywords, language whitelist. Articles killed before embedding, saving roughly 30% compute.

Gate 2 (s09). Corroboration rules after cosine subclustering, minimum 3 articles from 2 or more independent sources. **Syndication detection** via MinHash Jaccard (0.60 or higher) collapses wire service reprints into a single effective source. Bias risk guard flags subclusters where 80% or more of articles originate from a single source domain.

Gate 3 (s10). Multi scale substance validation, clusters formed at three geographic levels simultaneously (global, 17 regions, per country), each requiring minimum substance threshold. This ensures regional events are not drowned out by globally dominant stories.

Gate 4 (s11). Composite scoring with a 7 component formula (Table 2) determines hotspot promotion.

Table 2, Composite Scoring Components

Component	Weight	Basis
Volume	25%	log2(articles+1) divided by log2(pool_P95+1)
Recency	20%	Exponential decay, 12 hour half life
Diversity	20%	log2(unique_sources+1) divided by log2(20)
Topic	15%	Pillar based weight lookup
Velocity	10%	Lifecycle aware freshness ratio
T1 Convergence	+0 to 15%	Wire service coverage bonus
Anomaly Bonus	+0 to 20%	Z score deviation from baseline

3.1.5 Cross Day Lifecycle Management (s12)

Hotspots persist across pipeline runs through centroid based cross day matching (cosine 0.80 or higher AND shared country). A state machine tracks evolution. Born to active to fading (48h) to expired (72h) to revived. Trend signals (surging, escalating, stable, cooling, collapsing) are computed from inter run score deltas.

A **122 node domain classification** system assigns each hotspot to a domain tree node using embedding cosine, keyword validation, and pillar sanity checking, preventing misclassification of battlefield events as technology or similar errors.

3.1.6 LLM Significance Judge (s12b)

A dual provider LLM judge (Gemini + OpenAI in round robin) filters non significant hotspots. **Borderline events** (score = threshold) are consensus re judged by the other provider, both must agree for promotion. Domain reclassification suggestions are FK validated against the 122 node domain tree.

3.1.7 Dynamic Trust and Coverage Confidence (s14)

Source trust is computed empirically. `trust = lerp(static_trust, corroboration_rate, alpha)`, clamped to prevent extreme drift. Per region (17 regions) coverage confidence detects anomalies. Censorship (RSS drops more than 70%, API events remain), blackout (both drop more than 70%), surge (articles more than 3x baseline).

Hotspot output. Each promoted hotspot arrives at the forecasting pipeline as a structured record containing textual context (label, summary, top headlines), domain classification (primary + secondary pillars), signal metrics (z score, velocity, t1_convergence, anomaly bonus), geographic metadata (region, countries, coordinates), source articles with provenance, and lifecycle tracking (status, days active, velocity trends).

Input Quality Gate (6 checks). Before entering the forecasting pipeline, each hotspot must pass six additional filters. Minimum composite score (0.3 or higher), minimum article count, minimum source diversity, non empty summary, at least one named entity, and at least one country. Any failing hotspot is logged with a specific `SkipReason` and excluded.

3.2 Doctrine Selection

The system maintains a registry of 35 strategic doctrine entries organized into seven categories (Table 1), each with domain pillar affinity scores.

Table 1, Doctrine Categories

Category	Count	Primary Focus
Classical Statecraft	5	Historical strategic wisdom (Indian, Chinese traditions)
Geopolitics and Grand Strategy	12	Spatial power dynamics, alliance systems, containment
Hybrid and Cognitive Warfare	7	Information warfare, cyber operations, psychological operations
Economics and Trade	4	Financial instruments as strategic tools
Science and Technology	3	Disruption dynamics, AI governance, cyber security
Earth and Environment	3	Resource constraints, civilizational resilience
Infrastructure and Systems	1	Grid dependencies, technological momentum

Selection Algorithm.

Matrix lookup computes an affinity score for each doctrine based on the hotspot is primary and secondary domain pillars. Score formula is `(2 if HIGH_affinity else 1) x 10 + (1 if matches_primary_pillar else 0)`. Sort descending.

LLM routing passes the ranked candidate list to a specialized router agent, which selects the top 5 (`ROUTER_TOP_K`) with relevance scores and reasoning.

Quality filter retains doctrines with `relevance_score` of 0.5 or higher. If this yields fewer than 2 doctrines, the top 2 by score are retained regardless of threshold.

3.3 RAG Augmented Doctrine Analysis

For each selected doctrine, the system retrieves the top 20 semantically similar chunks from a per doctrine LlamaIndex VectorStoreIndex using local `all-MiniLM-L6-v2` embeddings (384 dimensions), reranks to the top 5 using a local `ms-marco-MiniLM-L-6-v2` cross encoder, deduplicates via Jaccard similarity (threshold 0.7) and truncates to 6,000 characters per doctrine, and then analyzes via a council agent that produces a structured `DoctrineAnalysis`.

The analysis contains a confidence score (0.0 to 1.0) indicating how applicable this doctrine is to the current situation. A direction (+1, 0, or negative 1) indicating whether the doctrine is assessment is escalatory, neutral, or de escalatory. And the key concepts, specific principles from the doctrine invoked in the analysis.

Dynamic Evidence Gathering. The council agent possesses a `retrieve_more` tool. During analysis, the LLM can autonomously request additional doctrine passages by specifying a `doctrine_id` and a targeted query. This is a **key agentic differentiator**. The system is not limited to a static context window but can actively seek more evidence when the pre fetched context is insufficient. Queries are validated against the set of selected doctrines, preventing hallucinated doctrine references.

Output validation. Council output is validated against expected doctrine IDs. Unknown IDs are dropped, missing doctrines are logged, and zero valid analyses triggers a hard failure, ensuring downstream synthesis never operates on fabricated analysis.

3.4 Question Generation

The system generates one question per resolution horizon band. **14 days**, **30 days**, and **90 days** (`RESOLUTION_HORIZON_BANDS`). Each forecast also includes three **time projections** within its output, at 1 month (30 days), 3 month (90 days), and 6 month (180 days) horizons, showing how the probability is expected to evolve over time. Each question must be binary (resolvable as true or false), include a base rate from historical reference classes, specify falsifiable resolution criteria, and pass a programmatic quality gate (detailed below).

Questions for horizons already covered by existing open forecasts (within plus or minus 3 days) are deduplicated and skipped.

Self Improving Question Quality Gate. Each question passes through a validator with a blocklist of **36 banned vague phrases** before reaching the forecaster. The blocklist is structured in three tiers.

Expert curated (design time) includes terms like tensions increase, situation worsens, and major development.

Criteria analyzer additions (Phase 4) are patterns surfaced by automated analysis of resolver ambiguity data, for example further developments, evolving situation, and widespread speculation.

L1 feedback loop (production) includes phrases that *passed* the blocklist but still caused AMBIGUOUS resolutions in production, such as according to reports, if confirmed by officials, and significant military activity.

This three tier evolution provides **empirical evidence** that the self calibrating feedback loop operates not only on forecast probabilities (via Brier decomposition) but on the quality of input questions themselves. The system autonomously improves the precision of what it forecasts.

3.5 Tool Augmented Bayesian Synthesis

The forecaster agent has access to seven deterministic mathematical tools.

Tool 1, Bayesian Update

Standard single step Bayesian update.

$$P(H|E) = \frac{P(H) \times LR}{P(H) \times LR + (1 - P(H))}$$

where $LR = P(E|H) / P(E|\neg H)$ is the likelihood ratio. Output is clamped to [0.01, 0.99].

Tool 2, Laplace Smoothed Base Rate

For small reference class samples (n less than 20), applies Laplace or Jeffreys smoothing.

Method	Formula
Laplace	$(s + 1) / (n + 2)$
Jeffreys	$(s + 0.5) / (n + 1)$

This prevents extreme priors from tiny datasets.

Tool 3, Temporal Projection (CAMEO to Goldstein to Weibull)

A chained calculation that converts event streams into time horizon probability projections.

Goldstein Escalation Index. Each event is classified into one of 20 CAMEO categories, each carrying a Goldstein scale weight ranging from +7.4 (aid provision) to negative 10.0 (mass violence). Events are weighted by recency, $w_i = e^{-\alpha \cdot d_i}$ where d_i is days ago and $\alpha = 0.05$.

Weibull Shape Derivation. The normalized escalation index (negative 1 to +1) maps to a Weibull shape parameter k via a dead band threshold. Score less than negative 0.3 (escalating) yields k = 1.5 + |score| x 1.5, clamped to max 3.0 (increasing hazard). Score greater than +0.3 (de escalating) yields k = 0.5 + (1 minus score) x 0.3, clamped to min 0.3 (decreasing hazard). The neutral band [negative 0.3, +0.3] yields k = 1.0 (constant hazard, exponential).

Survival Analysis. For each projection horizon, probability is computed as exponential $P(t) = 1 - e^{-\lambda t}$ where $\lambda = -\ln(1 - P_{total})/T$, or as Weibull $P(t) = 1 - e^{-(t/\eta)^k}$ where $\eta = T/(-\ln(1 - P_{total}))^{1/k}$.

Tool 4, Time Horizon Decay

Direct survival analysis projection when the hazard model is known a priori (without CAMEO events). Supports both exponential and Weibull models.

Tool 5, Confidence Interval Estimation

Computes statistically valid confidence intervals for binomial proportions.

Method	Use Case
Wilson	Frequentist, good coverage for moderate n
Jeffreys	Bayesian, better for extreme proportions

Tool 6, Evidence Weight Score

Converts qualitative evidence into calibrated likelihood ratios.

Signal Type	Base LR
Signpost confirmed	5.0
Signpost emerging	2.0
News article	3.0

Effective LR = base_LR$^{\text{relevance_score}}$. Weight of Evidence (WoE) = ln (LR), classified as strong (above 1.0), moderate (above 0.5), weak (above 0.1), or negligible.

Tool 7, Fermi Decomposition

Decomposes complex questions into 2 to 7 sub questions.

Chain Type	Formula
Conjunctive (AND)	$P = \prod P_i$
Disjunctive (OR)	$P = 1 - \prod (1 - P_i)$

Includes sensitivity analysis showing how much each sub question contributes to the composite, and identifies the weakest link.

3.6 Pre Computed Anchors

Before the LLM runs, the system deterministically pre computes two probability anchors.

Signal Anchor. Converts hotspot signal metadata (z score, velocity percentile, t1 convergence, anomaly bonus) into a composite likelihood ratio (capped at 6.0), then applies a Bayesian update to the base rate. This provides a `signal_boosted_posterior` that the LLM must justify departing from.

Doctrine Anchor. Aggregates all council doctrine assessments via **log odds addition**.

$$\log\text{-posterior} = \log\frac{P}{1-P} + \sum_i w_i \cdot d_i \cdot c_i$$

where $d_i \in \{-1, +1\}$ is the doctrine direction, c_i is the confidence, and w_i is the performance weight (from calibration history, or a uniform scaling factor if unavailable).

3.7 LLM Output Constraints

The forecaster output is strictly validated by Pydantic. Probability must fall in [0.01, 0.99]. Confidence interval is validated via `model_validator` that lo is less than probability which is less than hi. Time projections must be exactly 3 (1 month, 3 month, 6 month), each with 2 to 3 key shifts. All fields are non optional (no null outputs permitted).

.

4. Quality Assurance, The Grounding Framework

4.1 Eight Point Validation

Every LLM forecast is evaluated against eight grounding checks.

Extreme probability flags predictions below 0.05 or above 0.95 (overconfidence marker).

Base rate drift flags when the absolute difference between probability and base_rate exceeds 0.50 (insufficient anchoring).

CI too narrow catches width less than 0.04 (false precision).

CI too wide catches width greater than 0.70 (uninformative).

Non monotonic projections flag time projections that zigzag rather than consistently trend.

Flat projections flag when standard deviation of projection probabilities is less than 0.02 (copy paste pattern).

Missing tool reference flags when calibration notes lack any reference to calculation tool outputs.

Self reported ungrounded catches when the LLM explicitly flags its own output as ungrounded (tool failures propagate this).

Grounding score = passed_checks divided by 8.

4.2 Auto Heal Pipeline

Recoverable failures trigger automatic corrections.

Ungrounded forecasts are dampened 30% toward base rate. **Narrow CI** is expanded to probability plus or minus 0.15. **Wide CI** is contracted to probability plus or minus 0.35. **Non monotonic projections** are reordered to nearest monotonic sequence (ascending or descending, minimizing L2 distance from original).

4.3 Retry and Rejection

If grounding score is less than 0.5, the forecast is retried with an explicit prompt listing failed checks. After retry, if grounding score remains less than 0.5, the forecast gets `status = REJECTED` (persisted but excluded from distribution).

4.4 Advisor Output Validation (Second Quality Gate)

Strategic recommendations produced by the advisor agent pass through a **separate 8 point validator**, creating a dual quality gate architecture unique to MASX AI.

Check	Failure Mode
Entity diversity	All recommendations target the same entity type
Recommendation depth	Recommendations under 15 words (generic)
Doctrine basis depth	Justifications under 8 words (hand waving)
Doctrine grounding	Citations of doctrines absent from the council analysis
Urgency probability coherence	Immediate urgency paired with low probability (less than 25%)
Key driver coverage	Recommendations ignoring the forecast is identified key drivers
Geographic coherence	Recommendations ignoring the hotspot is region and countries
Duplicate detection	Jaccard word overlap of 0.75 or higher between recommendations

Auto heal. Duplicates are dropped. Urgency mismatches are corrected to the probability inferred level. Quality score = (8 minus unique_flags) divided by 8.

The combination of the 8 point forecast grounding validator (Section 4.1) and the 8 point advisor validator constitutes a **16 check dual quality gate**. No known competing system applies this level of deterministic output validation.

.

5. Self Calibrating Feedback Loop Architecture

The self calibration system is MASX AI is most strategically significant innovation, creating a **compounding competitive advantage**. Every resolved forecast makes the system measurably more accurate.

5.1 Automated Resolution Pipeline

Past due forecasts are resolved through a multi stage automated sweep.

Stage 1, Multi Turn Web Search. For each due forecast, the system executes up to 5 search turns (`RESOLVER_MAX_SEARCH_TURNS`), each returning up to 5 results. Search queries are programmatically constructed from the forecast is structured data.

Turn	Query Composition
1	Raw forecast question
2	Question + resolution criteria keywords (first 200 chars)
3 and 4	Question + top key drivers
5	Reserved for additional criteria

URLs are deduplicated across turns. The search provider (Serper to Tavily fallback) is injected via `SearchProvider` Protocol interface.

Stage 2, Hallucination Proof URL Verification. Before the LLM is evidence citations are accepted, `_strip_hallucinated_urls()` cross references every cited URL against the set of URLs actually returned by the search provider. Fabricated URLs, a well documented LLM failure mode, are silently removed.

Stage 3, LLM Resolution Judgment. The resolver agent receives the forecast is question, resolution criteria, key drivers, disconfirming evidence, and all search results. It produces a structured verdict including an outcome (true, false, or null), corroboration_count, contradictory_evidence_found, evidence_summary, driver_hits, driver_misses, disconfirming_hits, signals_missed, and criteria_quality (sufficient, ambiguous, unfalsifiable, or vague).

Stage 4, Corroboration Gate. A deterministic override. If `corroboration_count` is less than `RESOLVER_MIN_CORROBORATION (2)` and the LLM claims a definitive outcome, the verdict is forced to `AMBIGUOUS` and `outcome` is set to null. This prevents single source resolution.

Stage 5, Verdict Mapping and Brier Score. Cleanly resolved forecasts receive an immediate Brier score, $(p-o)^2$. Ambiguous verdicts receive no Brier score and are excluded from calibration to prevent data contamination. Post mortem data (driver hits and misses, signals missed) flows into the calibration feedback channels.

Stage 6, Criteria Quality Analysis. A dedicated `criteria_analyzer` module aggregates verdicts to identify systemic patterns causing ambiguous resolutions, for example recurring phrases like no clear evidence or conflicting reports. Patterns appearing 2 or more times are flagged as `top_failure_patterns`, enabling iterative improvement of question generation to produce more falsifiable criteria.

Stage 7, Search Health Monitoring. The sweep tracks `search_health = resolved / (resolved + failed)`. Below 0.50, a structured warning is logged. The sweep continues regardless. Individual resolution failures never halt the pipeline.

5.2 Brier Decomposition

The standard Brier score decomposition is computed.

$$\text{Brier} = \frac{1}{N}\sum_k n_k(\bar{p}_k-\bar{o}_k)^2 - \frac{1}{N}\sum_k n_k(\bar{o}_k-\bar{o})^2 + \bar{o}(1-\bar{o})$$

Reliability — Resolution — Uncertainty

where forecasts are binned into 10 probability bins, $\bar{p}_k$ is the mean predicted probability in bin k, $\bar{o}_k$ is the hit rate in bin k, and $\bar{o}$ is the overall base rate. This decomposition diagnoses *what kind* of error the system is making. Reliability errors (overconfidence or underconfidence) are correctable through anchoring guidance. Resolution failures require better signal processing.

5.3 Three Channel Feedback Architecture

The calibration system feeds back through three independent channels simultaneously. This multi channel approach is unique among forecasting systems.

Channel 1, Mathematical Weight Adjustment (Doctrine Aggregation)

Brier scores are computed per doctrine agent and inverted to produce reliability weights, which feed back into the `multi_doctrine_log_odds()` function.

$$\log\text{-posterior} = \log\frac{P}{1-P} + \sum_i w_i \cdot d_i \cdot c_i$$

where w_i is the calibration derived weight from `doctrine_perf_store`. Doctrines with historically low Brier scores (better accuracy) receive higher weights. New doctrines without resolution history receive a uniform default weight. The weight computation and storage infrastructure is complete. Weight injection into the live forecaster pipeline is scheduled for activation once sufficient resolution volume is reached.

Channel 2, Domain Specific Prompt Guidance (LLM Behavioral Correction)

The `build_calibration_context()` function injects performance feedback directly into the forecaster prompt. This includes system wide Brier score and decomposition, domain specific performance relative to average (your domain forecasts are roughly Xpp WORSE than average), explicit behavioral guidance when calibration is poor (anchor more tightly to base rates), and positive reinforcement when calibration is strong (maintain current methodology). This channel is **fully operational** in the current production pipeline.

This acts as a behavioral corrective. Even if the underlying LLM model changes, the prompt carries forward the system is accumulated calibration lessons.

Channel 3, Past Signals Missed (Error Learning)

Resolution post mortems record which key drivers and disconfirming signals the system missed. These `past_signals_missed` are injected into subsequent forecaster prompts for the same hotspot, creating an explicit learning mechanism. The resolution pipeline captures and persists all post mortem data. Injection into the forecaster prompt is scheduled for activation alongside Channel 1 as resolution volume grows.

5.4 The Compounding Effect

```
Day 1    Doctrines weighted equally (uniform prior)
Day 30   First resolutions, initial weights computed
Day 90   Weight divergence emerges (some doctrines 2x others)
Day 180  Pattern stabilizes, specific doctrines dominate their domains
Day 365  System has domain specific calibration profiles
```

This creates a **proprietary data asset**, hard won empirical knowledge about which strategic frameworks predict which domains most accurately, that cannot be replicated by a competitor starting from scratch.

5.5 Minimum Data Requirements

The calibration runner requires a minimum of 10 resolved forecasts (`MIN_FORECASTS_FOR_CALIBRATION`) before computing meaningful Brier decomposition. Below this threshold, the system operates with uniform weights and no prompt corrections. This is an honest limitation. The feedback loop is value is proportional to the volume of resolved forecasts.

.

6. System Resilience

6.1 Multi Provider Architecture

The system distributes LLM workload across two provider groups (Gemini 2.0 Flash and GPT 4o mini) in round robin, with cross provider fallback on transient failures. Each provider has independent rate limiting at 200 RPM.

6.2 Graceful Degradation

Every component degrades gracefully rather than failing.

Component	Failure Mode	Degradation
LLM provider	Rate limit or error	Fall through to secondary provider
Calculation tool	Invalid input	Return fallback guidance + flag ungrounded
RAG retrieval	LlamaIndex unavailable	Stub retriever returns empty results
Forecaster agent	All retries exhausted	Deterministic fallback forecast at base rate
Advisor agent	Exception	Forecast saved without recommendations
Resolution search	Search provider failure	Health metric tracked, sweep continues
Calibration store	Database error	Pipeline proceeds without calibration context

6.3 Deterministic Fallback Forecasts

When all synthesis attempts are exhausted, the system produces a fallback forecast. Probability is set to the base rate. CI is base_rate plus or minus 0.15. All 3 time projections sit at the base rate. The forecast is explicitly flagged as ungrounded.

This ensures the pipeline always produces output, with a transparent quality signal, rather than failing silently.

.

7. Discussion

7.1 Novelty and Competitive Positioning

The core innovation of MASX AI is the **integration** of twelve capabilities that no single competitor, or combination of competitors, replicates. Look, individual AI forecasting techniques are advancing rapidly. Metaculus AI bots are approaching superforecaster level Brier scores, and Recorded Future launched Autonomous Threat Operations in late 2025. But no system combines the full chain.

14 stage autonomous data pipeline. The data pipeline processes thousands of articles daily through 14 stages with 4 progressive quality gates (junk filter to corroboration to substance to composite scoring), producing only corroborated, multi source hotspots. No competitor applies multi gate corroboration with syndication collapsing and bias detection.

3 tier cross domain detection. A hybrid pipeline (static rules to embedding cosine to zero shot DeBERTa classifier) tags every article with cross domain relevance, all CPU only at $0 inference cost. No competitor runs comparable multi tier domain classification without GPU infrastructure.

Cross day lifecycle state machine. Hotspots are tracked across pipeline runs via centroid embedding matching (cosine of 0.80 or higher + country overlap), with a state machine (born to active to fading to expired to revived) and trend computation (surging, escalating, stable, cooling, collapsing). No competitor implements embedding based cross day event evolution tracking.

Diverse strategic corpus. 35 doctrines from multiple civilizational traditions provide genuinely diverse analytical perspectives, not just paraphrases of a single framework. No other automated system uses named strategic traditions as RAG grounded analytical agents.

Mathematical constraint. Seven calculation tools anchor LLM output to statistical reality. Tool references are validated post hoc by the grounding framework.

Automated quality assurance. The 8 point deterministic grounding validator catches and corrects systematic LLM biases before publication, unlike stochastic LLM reflection loops.

Self calibrating feedback loop. The three channel architecture (weight adjustment + prompt guidance + error learning) creates a compounding advantage absent from all known competitors. Good Judgment tracks human accuracy but does not feed it into Bayesian aggregation functions.

Fully autonomous pipeline. Signal ingestion through distribution, resolution, and calibration runs without human intervention. Metaculus AI bots forecast individual questions but do not generate questions, distribute content, resolve outcomes, or self calibrate. Competing platforms either augment human analysts (Palantir, Eurasia Group), require human forecasters (Good Judgment), or produce intelligence feeds rather than probabilistic forecasts (Recorded Future, Adarga).

Quantified decision support. Intervention windows (roughly X days until likely) and cost of inaction risk multipliers (waiting costs Yx more risk) translate probabilities into actionable timing advice. No competitor produces these outputs.

Dual quality gate with auto heal. The 8 point forecast grounding validator combined with the 8 point advisor output validator creates a 16 check deterministic quality enforcement

pipeline. Both gates auto heal recoverable failures rather than rejecting outright. No competing system validates and auto corrects both forecasts and recommendations.

Self improving question generation. The 36 term vague phrase blocklist evolves through three tiers (expert curated to criteria analyzer to L1 production feedback), demonstrating that the feedback loop improves not just forecast accuracy but input question quality.

Autonomous content production. The system operates as a complete intelligence media platform. Deterministic newsletter builder (no LLM, 7 chart data visualization), LLM generated LinkedIn posts with rotating chart types, weekly flagship analyses, and a fully autonomous podcast pipeline (LLM script to TTS to MP3 to Supabase Storage). No competitor produces daily audio briefings from forecast data without human involvement.

7.2 Limitations

The calibration feedback loop requires a minimum of 10 resolved forecasts to produce meaningful Brier decomposition. The system is accuracy advantage is currently theoretical and must be empirically validated as resolution volume grows. This is an honest limitation and the single most important milestone ahead.

The 4 gate corroboration architecture in the data pipeline provides strong quality filtering, but forecast accuracy is ultimately bounded by the coverage and diversity of the source feed network (currently 200+ sources across 17 regions).

Some forecasts cannot be cleanly resolved as true or false. These lead to ambiguous verdicts that do not contribute to calibration, which slows the feedback loop.

Some source texts are historical. Their applicability to novel scenarios (for example, AI driven warfare) requires ongoing evaluation. I have seen the doctrine analyses surface genuinely useful insights on technology competition, but the older military treatises are less naturally suited to cyber or space domains.

7.3 Future Directions

Expand the doctrine corpus to include emerging frameworks in AI governance, climate security, and space competition. Implement Monte Carlo simulation tools for complex multi factor scenarios. Develop real time signal integration for continuous probability updating between pipeline runs. Build a public calibration dashboard for external verification of forecasting accuracy. Publish calibration track record data as the resolution volume reaches statistical significance. The live system is publicly accessible at https://forecast.masxai.com.

.

8. Conclusion

MASX AI demonstrates that autonomous, research grade geopolitical forecasting is achievable. The combination of multi agent deliberation, RAG grounded analysis, tool constrained Bayesian synthesis, and closed loop self calibration turns raw signal data into calibrated probability estimates with quantified decision support, all fully autonomously. The self calibrating feedback loop creates a compounding competitive advantage. As resolution volume grows, per doctrine performance weights increasingly optimize the Bayesian aggregation, progressively improving system accuracy in a way that cannot be replicated by competitors without equivalent historical data. The hard part was wiring everything together. The compounding part just takes time.

.

References

Tetlock, P. E., and Gardner, D. (2015). *Superforecasting, The Art and Science of Prediction*. Crown.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1 to 3.

Goldstein, J. S. (1992). A Conflict Cooperation Scale for WEIS Events Data. *Journal of Conflict Resolution*, 36(2), 369 to 385.

Mackinder, H. J. (1904). The Geographical Pivot of History. *The Geographical Journal*, 23(4), 421 to 437.

Tetlock, P. E. (2005). *Expert Political Judgment, How Good Is It? How Can We Know?* Princeton University Press.

Christensen, C. M. (1997). *The Innovator is Dilemma*. Harvard Business School Press.

Mearsheimer, J. J. (2001). *The Tragedy of Great Power Politics*. W. W. Norton.

Bostrom, N. (2014). *Superintelligence, Paths, Dangers, Strategies*. Oxford University Press.