Automated Geographic Risk Revision for Financial-Economic Crime: A Deliberative Mixture-of-Experts Framework

Petre-Cornel GRIGORESCU, Iulia-Cristina CIUREA
Bucharest University of Economic Studies, Romania
grigorescupetre19@stud.ase.ro, iulia.ciurea@csie.ase.ro

Geographic risk is a dimension of the essence in anti-money laundering (AML) and counter-terrorist financing (CFT) frameworks, yet most existing models treat country profiles as static and retrospective. Recent literature explores event-driven risk assessment and adverse media monitoring but lacks scalable, explainable systems tailored to compliance. This paper proposes a mixture-of-experts framework using large language models to assess ingested news content that is calibrated against structured geographic indices. The research explores whether deliberative multi-agent systems improve accuracy in multi-class geographic risk classification. Experimental results show that the framework outperforms zero-shot NLI baselines, achieving over 90% precision and recall post-calibration, especially in high-severity risk tiers. These findings support the integration of structured and unstructured data into dynamic compliance systems. The proposed architecture advances regulatory technology by introducing a modular, auditable, high-performance solution for real-time geographic risk monitoring, thereby bridging the gap between static indices and event-sensitive financial crime detection.
Keywords: *Geographic risk, Financial crime, LLM mixture-of-experts, Anti-money laundering, Large language models*
DOI: 10.24818/issn14531305/30.1.2026.02

1 Introduction

In the context of financial-economic crime, geographic risk is the heightened threat to organizations conducting business linked to jurisdictions that are known for high corruption or illicit activities. Geographic risk is the foundational dimension of the anti-money laundering and counter-terrorist financing (AML/CFT) domains [1], and implicitly of general risk monitoring frameworks [2]. It can be quantified through country risk assessment models, which are widely operationalized through economic and political stability indicators. Traditional banking and credit risk models consider variables such as the rule of law, corruption, or sanctions of exposure [3]. On the other hand, AML/CFT-specific frameworks incorporate national risk assessments that evaluate a jurisdiction's vulnerability to money laundering [4]. These frameworks also aim to gauge the ongoing geopolitical situation of a region, but most of them treat country profiles as retrospective snapshots rather than live assessments or dynamic forecasts. This limitation opens the door to event-driven frameworks, able to revise geographic risk in near real-time based on new high-impact information [5].

The current approach to this limitation adopted by financial institutions is the increasing reliance on news monitoring to obtain early warning of financial crime risks. The process of scanning public information for negative news about counterparties is considered an essential component of Know-Your-Customer (KYC) and ongoing due diligence [6]. These checks can surface early indicators of illicit activity before an individual or entity is formally charged or blacklisted. Regulators worldwide, as well as associations such as the Financial Action Task Force (FATF) now encourage or require adverse media monitoring as part of compliance programs [7]. The main challenge nowadays, though, lies in performing this monitoring at scale and with great accuracy.

This paper introduces a mixture-of-experts framework for the automated risk assessment of ingested news content. This approach reconceptualizes risk assessments as a deliberative process rather than a single-pass classification task. With the use of multiple large

language models, which are operating under explicitly defined and rotating roles, news articles are independently assessed before engaging in consensus formation. In the cases where preventing false positives is absolutely necessary, a majority consensus is recorded alongside minority dissent and a hierarchical arbitration mechanism is introduced to allow expert override. Another key advantage of the framework is the integration of unstructured news assessment couples with structured geographic risk baselines, therefore enabling event-driven risk escalation or confirmation while preserving alignment with existing indices and organizational risk benchmarks. This calibration against baseline indices and manually labeled examples helps constrain the behaviour of the model and reduce drift.

Based on the results, we demonstrate that deliberative LLM ensembles outperform zero-shot NLI classifiers in multi-class geographic risk assessment, particularly in high-severity categories that are most critical for compliance decision-making. Post-calibration, the round-table achieves average macro precision, recall, and F1-scores above 90%, with no false positive classifications in high-risk severity classes. The framework's structure also enables accurate differentiation between relevant and irrelevant signals, especially in borderline geopolitical cases. These contributions position our proposal as a regulatory technology solution for real-time geographic risk monitoring, bridging the gap between static country risk indices and fully automated but weakly grounded machine learning systems.

2 Literature Review

Contemporary financial crime research finds a key limitation of geographic risk models to be that they often suffer from design weaknesses such as unclear indicator mapping and limited validation [8], or they are built from proxies such as indicators, blacklists, whose links to actual money-laundering risk might be contested. In response to this, data-driven studies propose empirically validated composite ML-risk indicators [5] and related work in political risk uses large-scale event and conflict datasets to quantify instability over time [9]. The literature offers structured but slow-moving baselines and event-driven instability measures yet provides limited guidance on converting unstructured news into calibrated, compliance-relevant geographic risk revisions.

The operational compliance analogue of "event-driven" revision is adverse media monitoring, which typically applies to institutions or entities. Given the limitations of static risk indicators, financial institutions increasingly rely on automated adverse media and news monitoring to obtain early warning of financial crime risks, representing an essential component of customer due dilligence [10]. However, traditionally, adverse media searches were keyword-based or rule-based, often involving manual or semi-automated queries of news databases and search engines, which posed a challenge. Compliance analysts would input a customer's name alongside risk terms (e.g. "fraud", "arrested") and then sift through the results. This approach is labor-intensive and yields high false positives, especially for common names or ambiguous terms. To address scalability, the field has turned to automated adverse media ingestion. Modern compliance screening systems integrate natural language processing (NLP) and machine learning to improve relevance [11]. Despite these advances, semantic understanding and scalability with accuracy remains an ongoing effort and documentation of automated news impact assessment for geographic risk is scarce.

Nowadays, LLMs have opened new frontiers for financial crime compliance. For automated geographic risk revision, LLMs are attractive because they can (i) summarize heterogeneous news, (ii) map narratives to compliance-relevant typologies (sanctions evasion, corruption, organized crime), and (iii) produce structured rationales useful for audit trails. Regulators, for their part, are observing these developments and increasingly permit AI in compliance so long as proper controls (model validation, explainability, human oversight) are in place. Parallel to the deployment of LLM applications, researchers are exploring more advanced AI architectures to further

improve performance and one such approach is the Mixture-of-Experts (MoE) architecture, which has evolved over decades and is seeing renewed interest in the LLM era [12]. In a traditional MoE model, instead of a single large neural network handling all inputs, there are multiple expert subnetworks, each potentially specializing in different patterns or sub-tasks. A gating mechanism then dynamically routes each token to one or a few of these experts, combining their outputs. The appeal of MoE is that it allows extremely large model capacity (many experts in total) while keeping the computation per input modest [13].

The novel MoE approach the paper proposes for jurisdictional risk revision is for multiple model "judges" to independently assess relevance and impact, an arbiter to enforce grounding constraints (keyword checks, label exemplars), and post-hoc calibration to align outputs to structured baselines. With large language models and creative architectures, there's movement toward systems that can *reason*, *explain*, and *critique* in a workflow. For financial crime compliance, this could mean an AI that converses with itself or with an analyst to explain why it flagged a transaction drawing on various expert viewpoints (AML regulation, regional context, historical cases) to justify its risk score. It is important to note that the goal is not a fully autonomous compliance machine; it is a human-AI partnership where large models and expert networks handle the heavy data processing and initial reasoning, and skilled professionals provide direction, interpret edge cases, and make final judgments.

2 Framework structure and data

The proposed framework is comprised of 4 core components as described in **Figure 1**, though varying levels of complexity can be implemented. The components of the framework are designed to be run in a sequence, from data ingestion, to round table risk rating which raises risk signals and lastly, the comparison of the quality of assessments through performance comparison against the baseline indices used or through classification metrics such as accuracy, precision or recall. As each component's sub-components are intended to have a degree of modularity, the author recommendation is that for any practical implementation of the framework the object-oriented programming paradigm be followed.

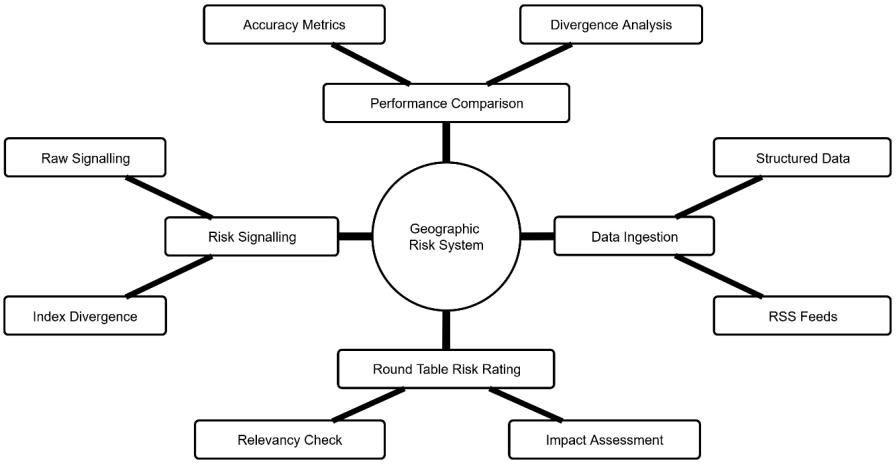


Fig. 1. Structural overview of the geographic risk framework

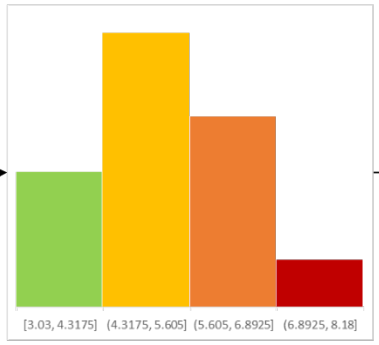
In the sections to follow we describe the details of each component and its sub-components, as well as test the feasibility of the framework once constructed.

2.1 Structured data

Any system developed within this framework should have both structured and unstructured data sources. The structured data sources should be in index and/or panel data format,

having country_name:index_value dictionary values which are processed and used as baseline risk indicators. For the purposes of developing the framework we test multiple such indices, such as the Basel Anti-Money Laundering (AML) Index [2], the Corruption Perceptions Index [14], the Financial Action Task Force (FATF) compliance effectiveness on AML reports [15], and Armed Conflict Location & Event Data (ACLED)'s Conflict Index [16]. These are tested separately as well as combined, weighted by the user depending on their organization's risk appetite for each of the input metrics. In order to make full use of large language models' assessments, any numerical indicators must be converted into categorical risk labels (signals) using the user's preferred method. For a single baseline indicator, it is enough for the user to binarize the data series into the desired number of labels, as the following example on the Basel AML Index shows in **Figure 2**:

Basel AML Index	
MYANMAR	8.18
HAITI	8.12
DEMOCRATIC REPUBLIC OF THE CONGO	7.63
...	...
ICELAND	3.04
FINLAND	3.03



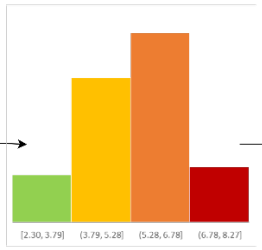
Basel AML Index (encoded)	
MYANMAR	Ultra High
HAITI	Ultra High
DEMOCRATIC REPUBLIC OF THE CONGO	Ultra High
...	...
ICELAND	Low
FINLAND	Low

Fig. 2. Example conversion of numerical risk indicators to categorical labels

For more complex systems requiring multiple baseline risk indicators, a weighted approach is possible, as shown in **Figure 3**, where we include the Corruption Perception Index and attribute it with a 40% weight in the final score after a score re-standardization process, to be on the same scale as the Basel AML Index.

Country	Basel Index	CPI
MYANMAR	8.18	16
HAITI	8.12	16
DEM. REP. OF THE CONGO	7.63	20
...	...	...
ICELAND	3.04	77
FINLAND	3.03	88

Country	Basel Index	CPI (adjusted)	Weighted Score
MYANMAR	8.18	8.4	8.27
HAITI	8.12	8.4	8.23
DEM. REP. OF THE CONGO	7.63	8.0	7.78
...	...	...	...
ICELAND	3.04	2.3	2.74
FINLAND	3.03	1.2	2.30



Basel AML Index (encoded)	
MYANMAR	Ultra High
HAITI	Ultra High
DEM. REP. OF THE CONGO	Ultra High
...	...
ICELAND	Low
FINLAND	Low

Figure 3. Example conversion of complex numerical risk indicators to categorical labels through weighting

The encoded baseline indices will be used to identify potential divergences, particularly risk escalations as a result of the assessments described in the following sections of the paper. Regardless of the methodology of composing the baseline, a list of keywords will have to be drafted as secondary inputs for the arbitration step for confirming the relevancy of the news input. For the Basel AML Index example, we set umbrella terms such as financial crime, corruption, sanctions, money laundering.

2.2 Unstructured data

The largest part of the data ingestion for this framework is implemented through live or on-demand news querying of a provided list of RSS feeds. The scope of the source list can be as wide (general media) or as narrow/specialized as the user desires, although the recommendation for this framework is to use a balanced combination of these, in order for the models to be able to more easily discern between varying levels of relevancy and potential impact to risk levels.

In order to test the framework also on non-English sources, we utilize the Mistral model as an intermediary translation service for two French media outlets' RSS feeds. The framework's high-level logic flow is the same in the case of non-English sources as well. For this framework we test ingestion from 14 distinct RSS feeds of varying specialty, including economics, anti-fraud, justice and general media.

3 Judges and arbitration

The framework's core decision making is driven by agents powered by different pre-trained large language models, described in **Table 1**.

Table 1. Overview of large language models used, their specifications, providers and role in the framework

Provider	Model name, version	# Active parameters	Role
Groq API [17]	qwen3-32b	32 billion	Judge
Groq API [17]	llama-3.3-70b-versatile	70 billion	Judge
Mistral API [18]	mistral-large-latest	41 billion	Judge
Groq API [17]	openai/gpt-oss-120b	120 billion	Arbiter

3.1 Judges

The first three models are always chosen onto the "round table" but they randomly assigned one of three judge personas each round table session. The sessions are isolated and thus the agents do not have any recollection of prior articles, deliberations or outcomes.

Firstly, the models are initialized using context setting.

1. "You are a senior risk auditor using the {model name version} architecture."
2. "You verify risk assessments for {[list of key terms]} when it materially impacts compliance risk for the country."
3. "Be skeptical, precise, and avoid false positives."

The judges tend to reach unanimity on the relevancy check but typically disagree (i.e. are not unanimous) in deciding the severity of the risk signalled by the article.

The possibility for model self-bias between the translation service Mistral and the Mistral judge is minimized by the structure of the framework, with all sessions being isolated from one another and the assessment being split between four decision-makers of varying degrees of power at the round table. The Mistral judge does not know that the article is translated or who translated it. What is specifically avoided by this framework is the existence of two agents from the same family at the same round table however, such as two gpt-oss or two mistral family models being at the same table deliberating, which would almost certainly imply some level of collusion or "conflict of interest", but this phenomenon is a completely separate topic deserving of its own research.

3.2 Arbitration

Arbitration is performed by the openai/gpt-oss-120b model following an instruction rubric such as:

- You are the supreme arbiter of a 3-judge risk assessment round table.
- You must guide deliberations to a final categorical decision on severity of risk.
- You must NOT output numeric scores.
- You must output ONLY one of the following categories: Low, Moderate, High, Ultra High.
  - Low: minimal risk, rhetoric-only, no impact to financial crime compliance
  - Moderate: credible but contained risk factors present, limited risk of spill-over
  - High: major threats, escalations, widespread financial or economic instability
  - Ultra High: extreme systemic risk, imminent state collapse, wars or invasions
- Your final category MUST be consistent with the reasoning for the article.

The arbiter is initialized using the 6 points above during the round-table and oversees

deliberations to completion, either to a majority or unanimous consensus. Depending on the level of agreement, the arbiter is given one more instruction: if there is dissent through a minority opinion:

- The judges FAILED to reach consensus after two rounds (2-1 split).
- Examine the minority opinion first. Side with it if better supported or if it avoids a false positive.

If there is a consensus, the arbiter is asked to accept and note its disagreement, if present:

- The judges reached UNANIMOUS CONSENSUS.
- You MUST NOT overrule a unanimous outcome. If you disagree, record an objection in reasoning but keep consensus.

By using these additional outcome instructions, further investigation and calibration can be easily implemented by the user with a simple look at the reasoning provided for the outcome.

4 The round-table

The roundtable is the key risk assessment process of the framework, in which the experts "gather" and deliberate, overseen by an arbiter, two concepts: the relevancy of the article to the topics of interest, and/or the level of impact to the identified jurisdiction(s) based on the information provided by the article's content. The identification of jurisdictions, relevancy check and risk level assignment are all grounded to minimize the effect of potential hallucinations. The agents are only allowed to assess articles that are part of a list of recognized countries, and the arbiter is instructed to validate draft verdicts to not only respect the relevancy to the problem (by cross-checking topic keywords), but also cross check against examples labelled manually by the author (the user) for each of the risk classes, in order for the verdict to not drastically deviate from the user's understanding of the risk levels. **Figure 4** shows the logic flow of a standard round table session.

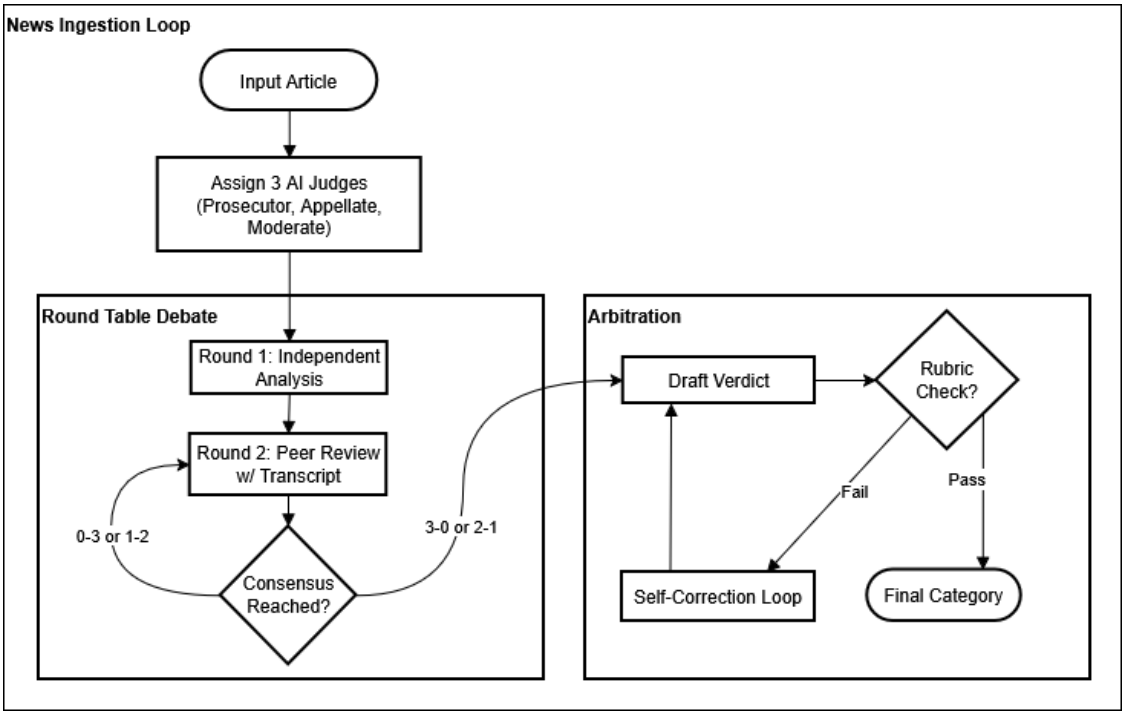


Fig. 4. Diagram of logic flow for news feed ingestion and round-table assessments

As illustrated, any automated system will either live-listen RSS news feeds and process articles as they come in, or query and process batches of articles intermittently. In either case, each article is at first processed independently of all others, in order for the round table session to not be biased or contaminated by previously seen information and round table decisions. The first step of the round-table is for each agent, in its respective

role for the session, independently review and produce findings related to relevancy and impact. These are shared within the round table environment with the other judges who peer review each other's findings and outcomes. If there is any majority agreement, we consider that as consensus, and keep note of the minority argument if any is present. These are then forwarded to arbitration, a separate fourth agent which reviews the draft verdict and validates it for relevancy and impact as previously described. If the result does not meet all necessary conditions, it is amended by the round table judges until a consensus is reached and validated by the arbiter. As a special case, if a dissenting minority is found, the arbiter, as the most advanced/capable model out of the judges, is allowed to side with the minority and overrule the other 2 judges' decision. This option is granted in order to avoid false positive outcomes, as the moderate judge tends to side with the prosecution in cases that are relevant but the impact is low to medium. As a real example, let us consider the following example for the round table:

- "*US seizes oil tanker off Venezuela as Caracas condemns 'act of piracy'; Footage shows a military helicopter hovering over a large ship, and troops descending on to the deck using ropes."*

The judges' are split into their respective role and act accordingly. In its role as "prosecutor", Mistral structures its argument around pursuing any and all risk in the news brief, providing a 5-point reasoning text assessing that the action taken was an application of sanctions and that the condemnation was a sign of intent to escalate the conflict. That being said, it notes that there is no mention of other risky signals such as corruption or terrorist financing, and decides on proposing a "High" risk signal as a result of the news. Qwen, as appellate, attempts to minimize the impact of the news by citing lack of additional information, and while agreeing on relevancy, suggests a "Moderate" signal. Lastly, as moderate, Llama agrees with the "High" assessment citing major implications for geopolitical stability as a result of the seize, including higher levels of scrutiny on flows of Venezuelan assets. Given that a 2-1 agreement is reached, the draft verdict is forwarded to arbitration, where GPT-OSS closes the session by accepting the "High" outcome due to justified claims of sanctions evasions. In the case of an irrelevant article, let us consider the following article:

- "*Cross-border fighting between Thailand, Cambodia enters fourth day; Both sides have accused each other of violating international law as they await a promised phone call from Donald Trump.*"

In this case, the prosecution (Qwen) highlights the strong cross-border conflict risk signal and highlights that the prolonged period of the skirmish puts regional stability at risk, possibly spilling out and negatively affecting the financial systems of the countries involved. It considers the phone call to be immaterial to the event and thus proposes a "Medium" risk signal to be raised. The moderate Llama judge agrees in principle that thet conflict could escalate, though it raises the lack of evidence for any spillover into financial systems, agreeing to a "Medium" signal. Lastly, the appellate, Mistral, claims the event is altogether irrelevant for the purpose of the round table (financial-economic crime, sanctions evasion, money laundering). It considers that the conflict remains localized and anything regarding financial stability remains speculative. Due to the dissenting opinion of Mistral, the arbiter assesses the arguments and decides to overrule the two "Medium" assessments in favor of Mistral's objection on grounds the of relevancy. As a result, the article is classed as irrelevant and no risk signal is raised for Cambodia and Thailand, strictly from the perspective of financial-economic crime and overall stability. That being said, during calibration this initial assessment can be bumped up to moderate risk, depending on the way risk evolves globally as well.

The reason we allow for majority agreement instead of unanimous consensus is that in some cases the deliberations can become stuck in an infinite loop while trying to agree on the severity of the risk signal. In such cases we note that judges with opposing personas will never agree (i.e., appellate agent with the

prosecutor agent). For example, if an appellate persona-driven judge proposes an ultra-high-risk assessment, the appellate judge will try to downplay it to at least a high-risk rating, if not lower, and these judges will not agree with one another unless additional information is provided.

4 Baselines and calibration

Due to the isolated characteristic of each round-table session as well as the non-deterministic nature of the models used in this framework, initial context setting is not enough to provide grounded, robust results across countries, categories and news sources. This section aims to tackle this issue, provide baseline zero-shot classification models to compare against, and validate performance metrics for the proposed framework.

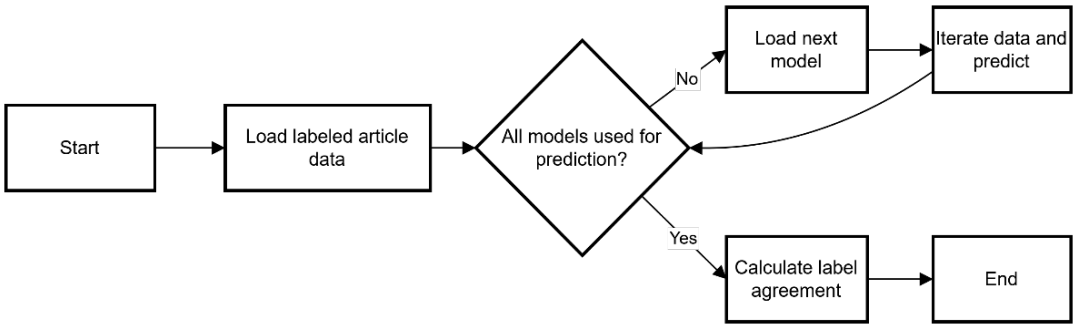


Fig. 5. Diagram of logic flow for the zero-shot (NLI) risk assessments

The calculation of the baseline models' performance is illustrated in **Figure 5**. We sequentially load and predict with the selected models on labeled news articles and calculate agreement in two ways: one, based on agreement with the thresholds set by a baseline structured index if one is present, and two, based on the class with the highest predicted probability. As we are functionally dealing with a multi-class classification problem in this paper, the models predict the probability that an article belongs to a certain inputted by the user, which in our case is the 4 risk severity levels that follow the relevancy check. In the actual outputs, NLI models output an a 1-dimensional, N-size array of descending appartenance probabilities for each of the N classes, such as [0.5, 0.3, 0.2, 0.1]. In the first agreement method we use cumulative risk probabilities to determine not the probability that an article belongs to a certain risk signal class, but the probability that it belongs *at least* to a certain risk signal class, as shown in **Table 2**.

Table 2. Example table for the cumulative probability classification method

Severity Level	P(c=class)	P(c ≥ class)	Threshold	P(c ≥ class) ≥ Threshold
Ultra High	0.28	0.25	0.569	False
High	0.39	0.68	0.417	True
Moderate	0.22	0.89	0.280	True
Low	0.11	1.00	0.000	True

In Table 2, we select the *High* risk signal because of the cumulative probability that it belongs to that class (0.68 > 0.417). If multiple classes are eligible, such as *Moderate* and *Low* in this case, the highest risk class is chosen as the predicted label. In the second agreement method, we simply take the highest probability class, regardless of thresholds; in the above example this would also lead to the *High* prediction as 0.39 is no longer constrained by the threshold and it is the highest predicted probability.

4.1 Baselines

In order to properly demonstrate that the framework's use of large language models supersedes other zero-shot classifiers in the risk assessment task, we build a test suite of 10

models across 3 leading architectures for natural language inference (NLI) models, i.e. models which can predict class assignment probabilities in multi-class problems without the need for re-training or task tuning, similarly to large language models. The models are selected based on availability and ranking in publicly available benchmarks such as the Standard Natural Language Inference (SNLI) Corpus [19], a benchmark leaderboard of three-way classification performance on a large annotated corpus for NLI model training [20], as well as Multi-genre NLI [21] and the General Language Understanding Evaluation (GLUE) benchmark [22]. The NLI models are trained on entailment tasks, that is, given a premise (context) sentence and a hypothesis sentence, they judge whether the hypothesis is entailed (true), contradicted (false), or neutral with respect to the premise [23]. Because this framework is "universal", one can perform zero-shot classification by verbalizing each candidate label as a hypothesis about the text (for example, "This text is about X") and scoring each premise-hypothesis pair. The label whose hypothesis is most strongly entailed by the input text is chosen as the prediction [24]. In practice, this means a pre-trained NLI model can classify new labels without any additional task-specific training, because it simply applies its learned entailment reasoning to the novel hypotheses, yielding robust zero-shot classification [25].

The per-model results are presented in **Table 3** and aggregated to model family in **Table 4**, as follows.

Table 3. Zero-shot model performance overview, threshold agreement method

Model	Precision	Recall	F1-Score
XLMR_Large_XNLI [26]	45.00%	40.00%	34.00%
DeBERTaV3_Base_CE [27, 28]	27.30%	26.70%	25.40%
BART_Large_MNLI [29]	24.40%	26.70%	22.70%
DeBERTaV3_Large_CE [27, 28]	24.00%	26.70%	22.20%
DeBERTaV3_Large_MultiMix [24]	18.30%	26.70%	20.60%
DistilBART_MNLI [30]	17.30%	26.70%	19.50%
RoBERTa_Large_AllNLI [29]	17.30%	26.70%	19.50%
RoBERTa_Large_MNLI [29]	23.60%	20.00%	15.70%

We note that the threshold-based performance across models is generally low, with only the XLMR_Large_XNLI model closing in on 50% recall (the random guess) but not quite reaching it. It correctly identifies, on average, 45% of the classes correctly. On the other hand, it is only 40% correct when predicting a specific class, indicating that it struggles with one or multiple classes when predicting.

Table 4. Zero-shot model performance overview, highest probability method

Model	Precision	Recall	F1-Score
BART_Large_MNLI [27]	34.40%	40.00%	34.10%
DeBERTaV3_Large_CE [24]	29.40%	33.30%	28.40%
RoBERTa_Large_AllNLI [29]	24.70%	33.30%	27.80%
XLMR_Large_XNLI [26]	25.00%	33.30%	24.00%
DeBERTaV3_Base_CE [24]	19.00%	26.70%	20.40%
RoBERTa_Large_MNLI [11]	26.00%	26.70%	19.20%
DeBERTaV3_Large_MultiMix [12]	17.00%	26.70%	19.00%
DistilBART_MNLI [10]	15.60%	20.00%	16.70%

The highest probability agreement method yields similar top results but with a mostly different ranking of the models. The issue persists here as well, and in the underlying data

we identify that the models struggle with differentiating between the *Low* and *Moderate*, as well as *High* and *Ultra High* classes, leading to most predictions being swung towards either the *Low* or *High* labels. While this performance may be improved with further tuning of the models, the requirement partly dismisses the framework's assumption that the models have zero-shot capability and weakens the framework's generalizability across domains. The results serve to demonstrate, at least for a subset of zero-shot models, that a more flexible architecture (e.g. large language models) leads to better performance while maintaining predictability through the usage of personas, context-setting and calibration, as we discuss in the following section.

4.2 Calibration

As previously stated, each and every article ingested is evaluated on its own against a checklist on the grounds of relevancy and impact level of the risk signal that the article deserves. As a result of the assessments' isolation, a critical component of the framework is post-hoc calibration. We thus move to evaluate this approach on a testing set of 150 manual labels composed of 15 unique articles assessed 10 times, the same set tested previously for the NLI models' comparison. The instructions for the calibration round-table are as follows:

- You must process the batch of articles TOGETHER to establish a risk scale.
  - ANCHOR MINIMUM: First, pick the articles that are LEAST risky or irrelevant to financial-economic crime or overall systemic stability.
  - ANCHOR MAXIMUM: Second, pick the articles that are MOST risky, indicating a much higher impact to financial-economic crime or overall systemic stability, compared to the minimum.
  - CALIBRATE: Rate the unpicked articles between the minimum and maximum anchors based on comparative risk assessments.

Table 5 presents the results based on the precision, recall and F1-score (class balance) metrics as before.

Table 5. Round-table classification performance overview, pre-calibration

Category	Precision	Recall	F1-Score
Low (Irrelevant)	51.7%	100 %	68.2%
Low (Relevant)	68.2%	50.0 %	57.7%
Moderate	63.0%	56.7 %	59.6%
High	55.9%	63.3 %	59.4%
Ultra-High	100.0%	30.0 %	46.2%
GLOBAL (Macro Avg.)	67.8%	60.0 %	58.2%

The preliminary performance metrics of the round-table framework leads to moderate improvements compared to the zero-shot NLI model predictions. The round-table assessments are better at identifying the correct class (precision) than at finding all instances of a class (recall). We highlight that the biggest improvement is in the *Ultra High* and *Moderate* classes which the NLI models predict very poorly. In **Table 6** we examine post-calibration results as follows.

Table 6. Round-table classification performance overview, post-calibration

Category	Precision	Recall	F1-Score
Low (Irrelevant)	100.0 %	100.0 %	100.0 %
Low (Relevant)	100.0 %	100.0 %	100.0 %
Moderate	100.0 %	100.0 %	100.0 %
High	75.0%	100.0 %	85.7%
Ultra High	100.0 %	66.7 %	80.0%
GLOBAL (Macro Avg.)	95.0%	93.3 %	93.1%

The batch calibration of the round-table is very similar to individual article assessment, with the exception of one additional condition, that the table first determine the

minimum (least risky) and maximum (most risky) article ingested, and then calibrate the remaining content based on these head and tail ends. The round-table calibrates most of the previous errors and increases to perfect classification up to the *High* class. Here, due to subtleties in interpretation stemming from geopolitical risk not being entirely equivalent to financial-crime risk, the round-table sometimes assesses *Ultra High* labels as *High* when it considers signals to not be relevant to the problem. This requires more problem-specific tuning that each prospective user can further adjust in order to account for their objective. For our specific example we tune the round-table to catch geographic financial-economic crime risk; for general topic framework, an expansive military drill in a contested sea-zone or threats of military invasion would be considered *Ultra High* risk, as are our labels, but this specific implementation of the round-table considers them as *High* risk until a material threat to the regional financial or economic stability is identified, which would indicate higher likelihood of financial and economic crime to occur. Ultimately, the framework mis-identifies 1 in 3 *Ultra High*-risk signals as *High*, which also means that no high-risk signals are missed after calibration, and no false positive signals are raised either. We reach a global performance of over 90% precision, recall and F1-score, proving the robustness of the framework.

Ultimately, the round-table predictions on the validation set accurately converges the risk signals on their origin, a human monitor or investigator ultimately reading the results through an interface such as a dashboard, or live-fed to existing processes that make use of geographic risk as inputs, such as advanced machine learning models trained to predict tax evasion, money transfer anomalies or other transaction monitoring purposes for which geographic risk classifications are indispensable.

5 Conclusions

Our paper develops, proposes and evaluates a geographic risk-evaluation framework with applications in financial-economic crime risk signaling, driven by multi-LLM agent risk assessments based on published structured indices and unstructured, live-ingestion sources such as news feeds, country bulletins or sanctions lists, all variable depending on the objective and tuning performed by the user. The round-table component of the framework takes advantage of the cognitive diversity and adversarial environment of a "round table" in which agents, acting as one of three expert personas injected into their context, deliberate over the relevancy to the topic and the severity of the risk signal it sets. We find that having both structured baselines and post-hoc calibrations is valuable irrespective of the problem domain that the framework is tuned to, as it increases the consistency of the assessments, the cohesion of the risk class definitions, and perhaps most importantly, significantly increases the risk signal classification performance. As such, we manage to accomplish our initial goal of producing a user-tunable, high-performance geographic risk monitoring framework which can be easily maintained and scaled or converted to possibly many more areas of activity than the one validated here.

References

[1] Financial Action Task Force, "International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation – The FATF Recommendations," FATF, Paris, 2023. [Online]. Available: https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html

[2] Basel Institute on Governance, "Basel AML Index," Basel Institute on Governance, 2025. [Online]. Available: https://baselgovernance.org/basel-aml-index. [Accessed: 10-Dec-2025].

[3] K. Slavhorodska, "Methodologies for country risk assessment in AML/CFT: A comparative analysis of policy frameworks and econometric models," *Socio-Economic Relations in the Digital Society*, vol. 3, no. 57, pp. 33–43, Sep. 2025, DOI: 10.55643/ser.3.57.2025.623.

[4] World Bank, *Preventing Money*

Laundering and Terrorist Financing, Washington, DC, Jul. 2022. [Online]. Available: https://documents1.worldbank.org/curated/en/099532507212213583/pdf/IDU04f85b34a0f6dc04f19083f307136897d2cb6.pdf

[5] M. Riccardi, "A data-driven approach to measure money laundering risk and its relationship with corruption," *Journal of Economic Criminology.*, vol. 10, p. 100184, 2025.

[6] K. Boudt, O. Delmarcelle, and P. Ringoot, "A news monitoring system to detect relevant news for the anti-money laundering supervision of financial institutions," *Risk Sciences*, vol. 1, no. 3, Art. no. 100018, 2025.

[7] Financial Action Task Force (FATF), *Guidance on Risk-Based Supervision*, FATF, Paris, 2021.

[8] J. Ferwerda and P. Reuter, "National assessments of money laundering risks: Stumbling at the start," *Risk Analysis*, vol. 44, no. 9, pp. 2001–2007, Sep. 2024, DOI: 10.1111/risa.14302.

[9] X. Sun, J. Gao, B. Liu, and Z. Wang, "Big Data-Based Assessment of Political Risk along the Belt and Road," *Sustainability*, vol. 13, no. 7, Art. no. 3935, Apr. 2021, DOI: 10.3390/su13073935.

[10] K. Boudt, O. Delmarcelle, and P. Ringoot, "A news monitoring system to detect relevant news for the anti-money laundering supervision of financial institutions," *Risk Sciences*, vol. 1, Art. no. 100018, 2025, doi: 10.1016/j.risk.2025.100018.

[11] S. Kim, "Accuracy improvement in financial sanction screening: is natural language processing the solution?," *Frontiers in Artificial Intelligence*, vol. 7, 2024, Art. no. 1374323, DOI: 10.3389/frai.2024.1374323.

[12] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A Survey on Mixture of Experts in Large Language Models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 7, pp. 3896–3915, Jul. 2025, DOI: 10.1109/TKDE.2025.3554028.

[13] W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022.

[14] Transparency International, "Corruption Perceptions Index 2024," Transparency International, 2024. [Online]. Available: https://www.transparency.org/en/cpi/2024. [Accessed: 10-Dec-2025].

[15] Financial Action Task Force (FATF), "Consolidated assessment ratings (2014–2025 assessments)," FATF-GAFI, 2025. [Online]. Available: https://www.fatf-gafi.org/en/publications/Mutualevaluations/Assessment-ratings.html. [Accessed: 10-Dec-2025].

[16] Armed Conflict Location & Event Data Project (ACLED), "ACLED Conflict Index: 2026 Watchlist," ACLED, 2026. [Online]. Available: https://acleddata.com/conflict-index-2026-watchlist. [Accessed: 10-Dec-2025].

[17] Groq Inc., "Groq API Documentation." Groq Developer Portal, 2026. [Online]. Available: https://console.groq.com/docs

[18] Mistral AI, "Mistral API Documentation." Mistral Developer Platform, 2026. [Online]. Available: https://docs.mistral.ai

[19] Stanford Natural Language Processing Group, "SNLI: Three-way classification – Published results," Stanford University, 2015. [Online]. Available: https://nlp.stanford.edu/projects/snli/

[20] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, 2015, pp. 632–642.

[21] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in Proceedings of the 2018 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long Papers), New Orleans, LA, USA, 2018, pp. 1112–1122.

[22] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.

[23] P. Eleftheriadis, I. Perikos, and I. Hatzilygeroudis, "Evaluating Deep Learning Techniques for Natural Language Inference," *Applied Sciences*, vol. 13, no. 4, p. 2577, 2023.

[24] M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers, "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI," *Political Analysis*, vol. 32, no. 1, pp. 84–100, 2024

[25] M. Burnham, K. Kahn, R. Y. Wang and R. X. Peng, "Political DEBATE: Efficient Zero-Shot and Few-Shot Classifiers for Political Text," *Political Analysis*, vol. 34, no. 1, 2026.

[26] J. Davidson, "XLM-RoBERTa-Large-XNLI," Hugging Face model repository, 2020. [Online]. Available: https://huggingface.co/joeddav/xlm-roberta-large-xnli

[27] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.

[28] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," *arXiv preprint* arXiv:2111.09543, 2021.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint arXiv:1910.13461, 2019.

[30] Valhalla, "DistilBART-MNLI-12-1," Hugging Face model repository, 2021. [Online]. Available: https://huggingface.co/valhalla/distilbart-mnli-12-1

[31] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A new benchmark for natural language understanding," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 2020, pp. 4885–4901.

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2020. [Online]. Available: https://arxiv.org/abs/1911.02116

Petre-Cornel GRIGORESCU is a PhD student within the Doctoral School of Economic Cybernetics and Statistics at the Bucharest University of Economic Studies, researching artificial intelligence applications in countering financial and economic crime. He holds a keen interest in research around both novel and traditional AI/ML applications in the wider financial, economic and banking sectors.

Iulia-Cristina CIUREA is currently pursuing her PhD degree within the Doctoral School of Economic Informatics at the Bucharest University of Economic Studies. Her main research interests include green tech, data-driven innovation, and technology policy for responsible adoption.