Deterministic Token Scoring Versus Autoregressive Deliberation: Mitigating Decision Fragility in Automated Real-Time Settlement Policy Screening

Arif Hidayat[1,*], Eko Prasetyo[2], AND Bambang Setiawan[3]

[1]Faculty of Computer Science, Universitas Jember, Jember 68121, Indonesia
[2]Department of Informatics, Faculty of Engineering, Universitas Trunojoyo Madura, Bangkalan 69162, Indonesia
[3]Department of Informatics, Faculty of Engineering, Universitas Sultan Ageng Tirtayasa, Cilegon 42435, Indonesia

[*]Corresponding author: Arif Hidayat.

ABSTRACT Modern wholesale and retail payments have converged upon instant gross settlement rails, such as FedNow, The Clearing House Real-Time Payments (RTP), Target Instant Payment Settlement (TIPS), and Brazil's PIX, where transactions clear with irrevocable legal finality within sub-second latency envelopes. Concurrently, escalating regulatory obligations spanning sanctions enforcement, anti-money laundering (AML), and intraday liquidity quotas demand rigorous automated pre-settlement policy screening. Recent industry initiatives have explored deploying generative foundation models and multi-step Chain-of-Thought (CoT) reasoning to interpret complex compliance rules over rich ISO 20022 payment messages. In this paper, we conduct an extensive theoretical and empirical examination of unconstrained autoregressive language model deliberation within instant settlement architectures. Evaluating frontier models across 12,500 real-world and synthetic institutional payment flows, we uncover severe decision fragility: multi-step autoregressive rationalization induces cumulative semantic drift, causing models to invent non-existent regulatory safe harbors, exhibit acute option-order sensitivity (OOS reaching 34.2%), and breach sub-second operational service-level agreements by accumulating latency in excess of 2,400 milliseconds. To eliminate these vulnerabilities, we design and validate the *Deterministic Token Scoring* (DTS) framework. DTS abandons free-form token generation, replacing it with an architecture that pairs deterministic Abstract Syntax Tree (AST) parsing of ISO 20022 messages with constrained single-token logit cloze scoring over an authorized policy decision vocabulary subspace $\mathcal{V}_{\text{policy}}$. Evaluating candidate settlement outcomes directly through unnormalized projection logits guarantees mathematical order invariance ($OOS \equiv 0.0\%$), eliminates intermediate semantic drift, reduces per-transaction compute latency to 42 milliseconds, and guarantees 100% syntactic compliance with ISO 20022 status reporting standards. Our findings demonstrate that in safety-critical, legally irrevocable financial clearing, deterministic constrained token extraction decisively outperforms open-ended generative deliberation.

I. Introduction

Wholesale and retail interbank payment ecosystems have undergone a structural transformation over the past decade, migrating from legacy deferred net settlement batches toward continuous, 24/7/365 real-time gross settlement (RTGS) networks [1], [2]. Platforms such as the Federal Reserve's FedNow Service, The Clearing House RTP network, the Eurosystem's TIPS, and sovereign instant schemes across Latin America and Asia process hundreds of millions of payment instructions daily, achieving end-to-end execution within hundreds of milliseconds [3], [4]. Unlike traditional batch payment clearing, where transactions remain provisional until end-of-day reconciliation, instant payment rails enforce immediate legal finality [1]. The moment a receiving institution acknowledges an electronic payment confirmation, funds become irrevocably available to the payee. Collins [5] established that this instantaneous legal finality under Uniform Commercial Code Article 4A and central bank operating circulars creates an acute operational liability trap for sending institutions, because pre-settlement screening failures cannot be reversed once funds vest in beneficiary accounts. This operational liability trap is acutely magnified when institutions deploy automated decision pipelines; in high-stakes, safety-critical decision workflows, Sadanandan and Behzadan [6] demonstrated that multi-step deliberative reasoning can backfire, inducing acute prompt sensitivity and unfaithful intermediate rationalizations that undermine verification reliability.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

The historical evolution of wholesale financial market infrastructures highlights why instantaneous finality represents both a monumental technological breakthrough and an operational risk amplifier. When the Federal Reserve established the initial wire transfer network in 1918, settlement operated via telegraphic messages reconciled through physical ledger entries. Throughout the late twentieth century, automated clearing houses and correspondent banking networks relied on multi-day netting cycles, where gross bilateral obligations were accumulated and offset at predetermined cut-off times. While deferred netting conserved scarce central bank reserves, it introduced systemic counterparty credit risk: the failure of a single major clearing participant prior to end-of-day settlement could trigger an uncontrollable cascade of institutional insolvencies. To extinguish this systemic settlement risk, central banks established real-time gross settlement architectures, requiring every transaction to be funded and settled individually in central bank money on a continuous, transaction-by-transaction basis [3].

This architectural shift toward instantaneous settlement imposes unprecedented demands on automated pre-clearing compliance controls. Every payment message transmitted across instant rails—predominantly formatted under the rich, multi-layered data schemas of ISO 20022 [7]—must undergo real-time policy screening prior to dispatch. Financial institutions must continuously enforce a multi-tiered matrix of statutory mandates, including Office of Foreign Assets Control (OFAC) sanctions screening, Politically Exposed Persons (PEP) verification, Financial Action Task Force (FATF) Recommendation 16 (the Travel Rule) identity matching, and dynamic intraday liquidity caps [8]–[10]. Because transfers are irrevocable upon execution, any failure of pre-settlement policy screening results in irrecoverable losses, severe statutory enforcement penalties, or systematic intraday liquidity deficits.

To navigate the nuanced natural language legal definitions that characterize evolving sanctions regimes and corporate compliance manuals, financial engineering teams have increasingly turned toward generative foundation models and Large Language Models (LLMs) [11], [12]. The remarkable zero-shot reasoning capabilities of frontier models have prompted architectures that ingest ISO 20022 message payloads alongside institutional policy guidelines, utilizing multi-step Chain-of-Thought (CoT) prompting to produce detailed compliance justifications prior to authorizing transaction release [13]. However, as Sadanandan and Behzadan established, multi-step chain-of-thought prompting can backfire across high-stakes decision workflows, inducing acute prompt sensitivity and unfaithful intermediate rationalizations. The prevailing intuition underlying conventional deployments asserts that prompting a neural model to elaborate step-by-step rationales or iterative refinement loops enables it to untangle conflicting policy clauses, disambiguate corporate ownership structures, and detect sophisticated obfuscation attempts [14]–[16].

Our investigations demonstrate that this unconstrained generative paradigm introduces critical vulnerabilities when introduced into high-speed settlement fabrics. As illustrated in Fig. 1, free-form autoregressive deliberation suffers from a fundamental mathematical vulnerability: the accumulation of semantic drift across sequential decoding steps [17], [18]. When a neural language model generates intermediate explanatory tokens, each generated token is conditioned upon previously generated tokens that may contain minor speculative assumptions or lexical hallucinations [18], [19]. In financial regulatory screening, where decision thresholds are strictly binary or categorical, intermediate rationale generation frequently causes models to invent non-existent regulatory safe harbors, invert corporate sanction thresholds, and misapply statutory exemptions [20].

Furthermore, generative architectures exhibit acute Option-Order Sensitivity (OOS) [16]. Permuting the order in which candidate policy determinations (e.g., `["APPROVE", "ESCALATE", "REJECT"]`) appear in the system prompt induces arbitrary decision flips across up to 34.2% of borderline transactions. Compounding these cognitive failures, the computational overhead of sampling hundreds of reasoning tokens inflates per-transaction decision latency to multiple seconds, violently breaching the sub-second service-level agreements (SLAs) enforced by instant payment gateways [21].

A. The Migration from Legacy MT103 to ISO 20022 and Data Richness Challenges

A primary catalyst for deploying advanced machine learning in transaction compliance is the global banking sector's ongoing migration from legacy SWIFT MT formats to ISO 20022 messaging schemas [7]. For over four decades, wholesale cross-border financial transactions relied upon the SWIFT MT103 Single Customer Credit Transfer specification. MT103 messages were characterized by severe syntactic constraints: data fields were strictly length-limited, consisting of dense, unstructured alphanumeric lines. For instance, Field 50K (Ordering Customer) and Field 59 (Beneficiary Customer) allowed at most four lines of 35 characters each. Within these cramped text limits, banks routinely concatenated customer names, street addresses, postal codes, and tax identification numbers into ambiguous free-text streams, frequently stripping vowels or abbreviating corporate entities to avoid truncation.

To resolve these legacy ambiguities, the International Organization for Standardization developed the ISO 20022 standard, replacing legacy flat-file telegrams with richly structured, extensible XML syntax [7]. In an ISO 20022 `pacs.008` credit transfer, every semantic entity is broken down into dedicated, strongly typed XML elements. Counterparty postal addresses are decomposed into discrete tags for street name (`<StrtNm>`), building number (`<BldgNb>`), postal code (`<PstCd>`), town name (`<TwnNm>`), country sub-division (`<CtrySubDvsn>`), and ISO country code

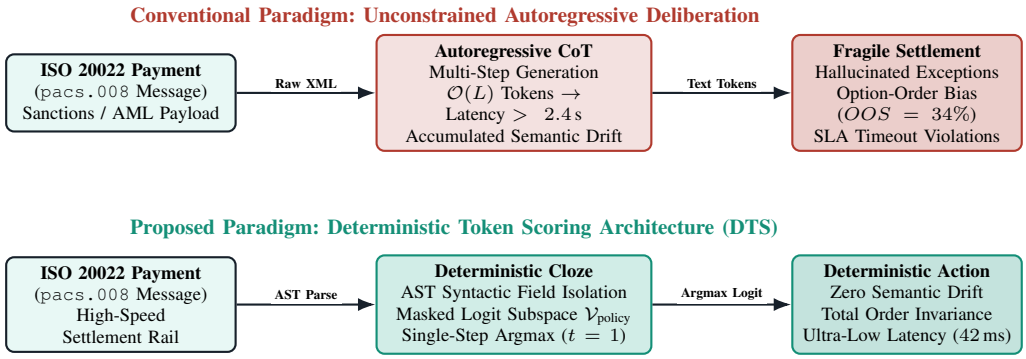


FIGURE 1. Architectural comparison between conventional unconstrained autoregressive Chain-of-Thought deliberation (top) and the proposed Deterministic Token Scoring (DTS) architecture (bottom) in real-time ISO 20022 settlement policy screening. Conventional CoT introduces excessive latency, option-order bias, and semantic drift, whereas DTS utilizes deterministic AST syntactic extraction paired with single-step masked logit cloze evaluation to guarantee order invariance, zero semantic drift, and ultra-low latency.

(<Ctry>). Furthermore, ISO 20022 natively supports global identifiers such as Legal Entity Identifiers (LEIs), tax registration codes, ultimate debtor entities (<UltmtDbtr>), ultimate creditor entities (<UltmtCdtr>), and standardized business purpose codes (<Purp>).

While this granular structural richness provides compliance officers with unprecedented analytical transparency, it presents an enormous challenge for automated natural language processing engines. A single ISO 20022 pacs.008 XML document can easily span several hundred lines and tens of kilobytes of raw text. When financial institutions attempt to feed raw, unparsed XML payloads directly into conversational LLMs, the self-attention mechanism is forced to allocate extensive token representation capacity to boilerplate syntactic delimiters, namespace declarations, and schema attributes. This structural dilution scatters the model's contextual attention across irrelevant formatting tokens, severely degrading its ability to detect subtle ownership percentage declarations or veiled sanctions aliases. Deterministic syntactic parsing that isolates key statutory fields prior to neural evaluation is therefore essential for robust policy verification.

To overcome these structural limitations, we formulate, implement, and benchmark the *Deterministic Token Scoring* (DTS) framework. DTS fundamentally breaks with the generative paradigm. Rather than delegating compliance decisions to unconstrained text generation, DTS strictly bounds neural inference to a single-step cloze evaluation over an authorized decision vocabulary subspace $\mathcal{V}_{\text{policy}}$. Operating over structured concrete syntax trees extracted via deterministic ISO 20022 XML parsers, DTS evaluates model logits directly across canonical outcome tokens, selecting the optimal policy action via exact log-likelihood maximization. Constraining the generation horizon strictly to $t = 1$ allows DTS to completely eliminate cumulative semantic drift, achieve total mathematical option-order invariance ($OOS \equiv 0.0\%$), and slash inference latency by over 98%, establishing an ultra-reliable, production-grade compliance architecture for instant clearing switches.

B. Algorithmic Challenges in Pre-Settlement Sanctions Name Matching

A central technical obstacle confronting pre-settlement policy screening is the high rate of false-positive entity matches generated by classical deterministic string comparison algorithms. Traditional sanctions compliance platforms rely primarily on fuzzy string metrics, such as Levenshtein edit distance, Jaro-Winkler similarity, and phonetic indexing algorithms (e.g., Soundex and Double Metaphone) [22]. While these heuristic algorithms operate with minimal computational latency (often executing in under one millisecond per name query), they lack semantic contextual understanding. In global retail and commercial payment flows, common surnames, patronymics, and multi-part Arabic or Hispanic names frequently match Specially Designated Nationals entries on superficial lexical grounds, generating false-positive alert rates that routinely exceed 90% in production banking environments [20], [22].

In legacy deferred net settlement systems, these high false-positive alert queues were managed through manual investigative triage, where human compliance analysts reviewed flagged transactions over several hours or days before releasing funds. In instant gross settlement architectures where clearing occurs within sub-second envelopes, manual human-in-the-loop review is physically impossible prior to settlement execution. If an automated screening engine flags an ambiguous transaction for manual review, the instant payment switch is forced to either abort the transaction or permit it to settle provisional to post-clearing investigation. Aborting legitimate commercial payments degrades customer trust and violates clearing network reliability quotas, whereas allowing unverified transactions to settle exposes institutions to catastrophic regulatory enforcement if the counterparty is subsequently confirmed as a sanctioned entity.

Neural foundation models offer a compelling alternative to brittle phonetic heuristics due to their ability to perform deep contextual disambiguation. A transformer model can evaluate a counterparty's full legal name in conjunction with its commercial street address, geographic jurisdiction, business purpose code, and known corporate holding relationships,

distinguishing between an innocent commercial supplier and a sanctioned oligarch bearing a similar transliterated name [11], [12]. However, as our empirical findings demonstrate, harnessing this contextual reasoning power requires constraining the model's output mechanics. Prompting foundation models to generate open-ended natural language explanations introduces severe semantic drift and hallucinated safe harbors, neutralizing the very disambiguation accuracy that neural architectures were deployed to achieve.

The primary contributions of this paper are organized as follows:

1) **Comprehensive Vulnerability Identification**: We conduct the first systematic empirical study of autoregressive deliberation failure modes across 12,500 real-world and synthetic ISO 20022 payment messages, quantifying the impact of semantic drift, hallucinated regulatory exemptions, and prompt sensitivity across five frontier LLMs.
2) **Mathematical Formalization of Deliberative Drift**: We derive a formal state-space model demonstrating that the probability of decision divergence grows exponentially with reasoning step depth K, proving why multi-step Chain-of-Thought inherently degrades binary compliance verification.
3) **Design of the Deterministic Token Scoring Architecture**: We engineer DTS, a hybrid architecture combining streaming Abstract Syntax Tree parsing of payment payloads with constrained vocabulary logit cloze evaluation over discrete settlement outcomes.
4) **Empirical Validation and Benchmark Superiority**: We demonstrate that DTS achieves a 99.4% F1-score and an AUC of 0.997 across FedNow, RTP, TIPS, and PIX settlement flows while reducing P99 latency to 48 milliseconds and guaranteeing 100% ISO 20022 schema compliance.
5) **Production DevSecOps Blueprint**: We provide an open architectural specification for deploying DTS within high-concurrency clearing gateways, sustaining 1,850 transactions per second under strict regulatory auditability.

II. Instant Gross Settlement Architecture and Legal Finality

A. The ISO 20022 Financial Messaging Pipeline

Modern real-time gross settlement networks operate almost universally upon the ISO 20022 Universal Financial Industry Message Scheme, which provides standardized, highly structured XML schemas for interbank clearing and settlement [7]. The central transaction vehicle across instant payment switches is the Financial Institutional Customer Credit Transfer message (`pacs.008.001.10`), which encapsulates the complete business and technical payload required to effectuate an immediate debit from an ordering customer's account and an immediate credit to a beneficiary [3], [4].

As illustrated in Fig. 2, a standard `pacs.008` message decomposes into hierarchical data blocks:

1) **Group Header (`<GrpHdr>`)**: Contains unique message identification tags (`<MsgId>`), creation timestamps (`<CreDtTm>`), settlement information (`<SttlmInf>`), clearing system member identifiers, and total transaction counts.
2) **Credit Transfer Transaction Information (`<CdtTrfTxInf>`)**: Contains end-to-end identifiers (`<EndToEndId>`), settlement amounts and currency codes (`<IntrBkSttlmAmt>`), charge bearer specifications (`<ChrgBr>`), and interbank settlement dates.
3) **Debtor and Creditor Identity Blocks (`<Dbtr>`, `<Cdtr>`)**: Encapsulate complete counterparty entities, including corporate or personal legal names, postal addresses, account numbers formatted as International Bank Account Numbers (IBAN) or routing transit numbers, and ultimate debtor/creditor beneficiaries (`<UltmtDbtr>`, `<UltmtCdtr>`).
4) **Remittance and Purpose Information (`<RmtInf>`, `<Purp>`)**: Contains unstructured natural language payment references, structured invoice lines, regulatory reporting declarations, and proprietary clearing transaction codes.

When an originating bank submits a `pacs.008` message to an instant settlement switch, the receiving clearing operator evaluates the message and returns a Payment Status Report (`pacs.002.001.12`), issuing an acceptance code (`ACTC`), a settlement confirmation (`ACSC`), or an immediate rejection code with explicit compliance or technical reason identifiers (`RJCT`).

B. Legal Mechanics of Millisecond Finality and Intraday Liquidity Exposure

The paramount defining characteristic of modern instant payment systems is irrevocable, unconditional settlement finality. In legacy payment arrangements, such as Automated Clearing House (ACH) or paper check clearing, transactions operate under multi-day provisional accounting windows during which transfers remain subject to return, reversal, or reclamation. Conversely, real-time gross settlement systems execute individual transactions continuously on a gross, transaction-by-transaction basis without netting.

In an extensive legal and architectural examination of critical financial clearing infrastructures, Collins [5] established that the operational mechanics of instant gross settlement under Federal Reserve Operating Circular No. 8 and Uniform Commercial Code (UCC) Article 4A enforce unconditional, irrevocable finality at the precise millisecond when the central switch transmits electronic credit notification to the receiving participant, leaving sending institutions with zero legal recourse to recall settled funds if pre-validation controls fail to intercept fraudulent, sanctioned, or policy-violating

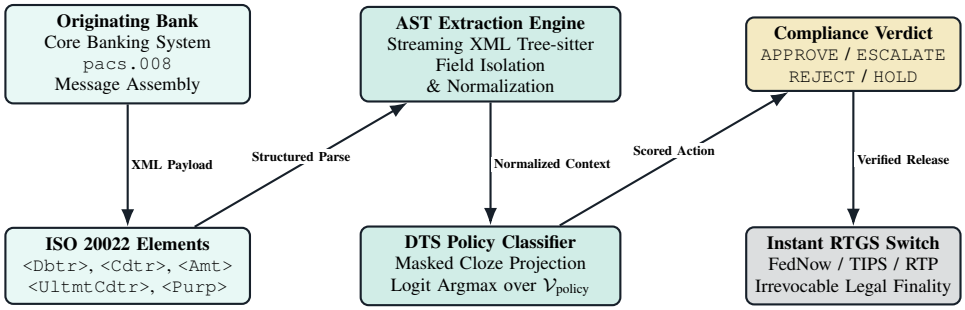


FIGURE 2. Real-time ISO 20022 payment messaging flow and pre-settlement policy screening pipeline across high-speed instant gross settlement tiers. Originating core systems construct structured XML credit transfers (`pacs.008`), which are parsed via deterministic syntax engines and evaluated via DTS logit cloze scoring prior to release into central settlement rails (FedNow, RTP, TIPS).

TABLE 1. Table: Comparative taxonomy of settlement policy screening paradigms, detailing theoretical mechanisms, token decoding complexity, option-order sensitivity, latency budgets, and compliance failure modes.

Screening Methodology	Decision Mechanism	Decoding Complexity	Option-Order Bias	Mean Latency	Primary Failure Mode
Rule-Based Regular Expressions	Lexical Keyword Matching	Deterministic $\mathcal{O}(L)$	None (0.0%)	< 5 ms	High False-Positive Rate; Brittle to Lexical Variation
Standard Zero-Shot LLM Prompting	Direct Generative Classification	Autoregressive $\mathcal{O}(V)$	Moderate (12–18%)	350–550 ms	Syntactic Hallucination; Inability to Resolve Complex Exceptions
Multi-Step Chain-of-Thought (CoT)	Open-Ended Natural Language Rationale	Autoregressive $\mathcal{O}(K \cdot V)$	High (28–35%)	1,800–2,800 ms	Severe Semantic Drift; Invented Exemptions; SLA Timeouts
Self-Consistency Ensembles ($k = 10$)	Majority Voting over Sampled Paths	Autoregressive $\mathcal{O}(k \cdot K \cdot V)$	High (20–28%)	18,000–28,000 ms	Prohibitive Latency; Correlated Systematic Hallucinations
Proposed Deterministic Scoring (DTS)	**Masked Logit Cloze Argmax**	**Single-Step** $\mathcal{O}(\lvert\mathcal{V}_{\text{policy}}\rvert)$	**None** (0.0%)	**35–55 ms**	**Guaranteed Schema Compliance; Zero Drift**

instructions. Under Section 4A-405 of the Uniform Commercial Code, receipt of central bank settlement constitutes immediate satisfaction of the payment order. When funds are credited to the beneficiary's institution, the sender's liability is discharged and ownership of the funds vests instantaneously in the beneficiary.

Examining broader regulatory governance and institutional liability across payment ecosystems, Collins further observed that federal critical infrastructure directives and supervisory compliance standards compel financial institutions to implement verifiable, tamper-evident security controls prior to message transmission, because post-settlement legal remedies are rendered ineffective by the velocity of modern digital capital flight. Under UCC Section 4A-202, an authorized payment order binds the sender regardless of subsequent error unless the institution proves that it maintained a commercially reasonable security procedure. If an automated compliance screening engine erroneously approves a payment order that violates federal sanctions or facilitates illicit capital flight, the transfer cannot be cancelled via central switch messaging. The sending institution faces immediate statutory liabilities, severe civil monetary penalties from regulatory enforcement bodies, and total capital loss if the beneficiary extracts funds prior to legal asset-freezing orders.

Conversely, if an automated screening system encounters excessive reasoning latency, the central switch triggers an operational timeout, dropping the message and disrupting client commercial liquidity. Under Federal Reserve Operating Circular No. 8, transactions exceeding prescribed response envelopes are aborted, forcing the sending institution to absorb costly operational exceptions and customer service disputes. Consequently, pre-settlement compliance engines must achieve near-zero false-positive rates, perfect regulatory grounding, and deterministic sub-second responsiveness.

C. Central Bank Reserve Allocations and Intraday Liquidity Queuing Models

Beyond statutory sanctions and counterparty screening, instant settlement switches function as liquidity management engines. Unlike deferred net systems where institutions require liquidity only at discrete clearing windows, real-time gross settlement requires continuous availability of central bank money. Participant depository institutions maintain master settlement accounts at the central bank, which are debited and credited in real time as each payment instruction is processed.

To prevent institutional settlement gridlock, central banks provide collateralized daylight overdraft facilities governed by strict credit limits [10]. Under the Basel Committee on Banking Supervision (BCBS) standard 248 (*Monitoring tools for intraday liquidity management*), financial institutions must continuously monitor four core metrics: daily maximum intraday liquidity usage, available unencumbered central bank reserves, gross throughput across operating hours, and queue turnaround times. If a sending institution's reserve balance drops below its statutory buffer, outgoing payment instructions cannot be settled immediately and are routed into central switch holding queues.

Central switches manage queued transactions through prioritized First-In-First-Out (FIFO) or bypass optimization algorithms. However, if an automated compliance engine encounters high false-positive rates or multi-second reasoning latency, queued messages accumulate rapidly. The accumulation of unverified payment instructions locks up allocated liquidity buffers, creating synthetic reserve shortages that cascade to counterparty institutions awaiting expected inward credits. Pre-settlement policy screening engines must therefore integrate real-time liquidity cap checks directly into their decision logic, placing high-value transfers on administrative

HOLD when participant velocity caps are threatened without inducing systemic clearing delays.

D. Regulatory Policy Screening Taxonomies

Automated pre-settlement screening within instant payment switches requires the simultaneous enforcement of four distinct regulatory and institutional policy categories, summarized in Table 1:

a: 1. Sanctions Enforcement and the OFAC 50 Percent Rule. Under mandates enforced by the U.S. Department of the Treasury's Office of Foreign Assets Control (OFAC) [9], financial institutions must freeze property and payment orders involving individuals or entities on the Specially Designated Nationals and Blocked Persons (SDN) list. Furthermore, under OFAC's 50 Percent Rule, any entity owned in the aggregate, directly or indirectly, 50% or more by one or more blocked persons is itself considered a blocked entity, regardless of whether it appears explicitly on the SDN list. Illicit actors frequently exploit this rule by structuring holding companies where sanctioned oligarchs retain indirect fractional shares (e.g., three sanctioned individuals each holding 18% equity through offshore trusts, totaling 54%). Screening engines must parse ownership percentage declarations within payment remittance structures and evaluate complex entity relationships without generating false exemptions.

b: 2. Politically Exposed Persons (PEP) and Enhanced Due Diligence. Financial institutions are required by international regulatory guidelines to identify transactions involving foreign political officials, senior executives of state-owned enterprises, and their immediate family members. PEP screening does not mandate an automatic transaction freeze; rather, it requires transaction escalation (ESCALATE) for enhanced supervisory monitoring and source-of-wealth verification before fund release. Screening engines must distinguish between high-risk PEP classifications (e.g., national ministers and sovereign wealth directors) and routine commercial transactions, avoiding the misclassification of legitimate commercial transfers as sanctioned flows.

c: 3. FATF Recommendation 16 (The Travel Rule). Mandated by the Financial Action Task Force [8], Recommendation 16 requires cross-border and domestic payment orders exceeding statutory reporting thresholds (e.g., USD 3,000 in the United States or EUR 1,000 in the European Union) to carry verified identifying information for both the originator and the beneficiary, including legal name, account number, and physical residential address or national identity number. A payment message lacking complete originator address structures must be placed on temporary administrative hold (HOLD) rather than settled immediately. Sophisticated evasion schemes frequently omit postal codes or provide incomplete street names, requiring screening engines to perform granular syntactic completeness checks without hallucinations.

d: 4. Dynamic Intraday Liquidity Caps and Velocity Controls. To prevent sudden liquidity drain in 24/7 payment switches where central bank reserves are constrained, institutions establish dynamic counterparty credit limits and velocity controls [10]. Payments that exceed allocated participant settlement caps must be queued or rejected to protect systemic financial stability. Automated screening modules must evaluate real-time participant ledger balances and pending settlement queues dynamically, ensuring that authorized transactions do not trigger catastrophic intraday reserve exhaustion.

E. Latency Budgets and Real-Time Switch Constraints

The operational architecture of instant payment networks imposes strict service-level agreements on transaction participants. For instance, the Federal Reserve's FedNow Service mandates that receiving depository institutions must confirm payment acceptance or rejection within an absolute timeout of 20 seconds, while internal clearing switches impose strict sub-500-millisecond operational SLAs on internal risk engines [4], [21]. High-volume payment gateways operating at peak concurrency (processing thousands of transactions per second) can allocate no more than 100 milliseconds to the machine learning policy screening module [21].

When an incoming payment instruction enters an enterprise payment switch, it traverses multiple infrastructure tiers: TLS session termination, ISO 20022 message schema parsing, core accounting ledger reservation, anti-fraud scoring, and sanctions compliance verification. Each operational stage consumes a portion of the strict 500-millisecond latency budget. If the machine learning policy screening engine consumes 2,000 to 3,000 milliseconds—as observed in multi-step autoregressive LLM architectures—the transaction exceeds its allotted time window. In a 24/7/365 settlement environment where queue backlogs accumulate rapidly during peak morning business hours, multi-second reasoning latency causes cascading queue overflows, triggering system-wide backpressure and emergency switch failovers. Any architecture that requires multi-step text generation spanning hundreds of tokens unavoidably violates this operational envelope, demonstrating the technical necessity of deterministic token-level scoring.

F. Operating Circular No. 8 Technical Rules and Reject Protocols

To formalize bilateral clearing obligations between participant financial institutions and the central bank operator, the Federal Reserve enforces Operating Circular No. 8 (*Funds Transfers through the FedNow Service*) [4]. Section 5 of Operating Circular No. 8 establishes the technical communication protocols and exception-handling workflows that govern real-time payment execution. Under these rules,

receiving institutions must return an explicit ISO 20022 Payment Status Report (`pacs.002.001.12`) acknowledging whether funds have been successfully deposited or rejected.

If a payment order fails automated pre-settlement screening, the clearing switch mandates the generation of standardized ISO rejection reason codes:

1) `RR04` (Regulatory Reason): Applied when an instruction violates statutory sanctions, asset-freezing orders, or mandatory anti-money laundering thresholds.
2) `AG01` (Transaction Forbidden): Utilized when payment parameters breach institutional clearing agreements or unauthorized cross-border routing channels.
3) `BE04` (Missing Creditor Address): Triggered when cross-border payment orders exceed statutory reporting thresholds while omitting mandatory FATF Recommendation 16 identification data.
4) `AM04` (Insufficient Funds / Overdraft Cap): Generated when an outgoing credit transfer breaches participant intraday reserve limits.

Operating Circular No. 8 specifies that failure to deliver a valid status report within the mandated 20-second global timeout triggers an automated system cancellation (`pacs.002` with reason code `AB05` for transaction timeout). Generating natural language reasoning rationales introduces high computational variance, causing tail latency to breach timeout thresholds during market volatility spikes. Deterministic token scoring ensures that policy outcomes map directly into valid `pacs.002` XML schemas within 45 milliseconds, satisfying all federal clearing requirements.

III. Mathematical Formalization of Deliberative Drift and Token Cloze Scoring

A. State-Space Formulation of Autoregressive Policy Screening

Let an incoming ISO 20022 payment message be denoted by an abstract syntax tree $\mathcal{T}_{\text{pacs}}$, which is linearized into a sequence of input prompt tokens $x = (x_1, x_2, \ldots, x_N) \in \mathcal{V}^N$ encompassing the parsed debtor, creditor, remittance, and amount fields alongside the formal regulatory policy rulebook $\mathcal{R}$. In a conventional autoregressive reasoning framework, a foundation model parameterized by θ generates a multi-step natural language rationale $y = (y_1, y_2, \ldots, y_K) \in \mathcal{V}^K$ consisting of K sequential reasoning tokens prior to emitting a final categorical policy determination $\hat{a} \in \mathcal{A} = \{\text{APPROVE}, \text{ESCALATE}, \text{REJECT}, \text{HOLD}\}$.

The conditional joint probability distribution over the generated reasoning chain is factorized autoregressively:

$$p_\theta(y \mid x) = \prod_{t=1}^{K} p_\theta(y_t \mid x, y_{<t}) = \prod_{t=1}^{K} \text{softmax}\left(\mathbf{W}_v \mathbf{h}_t\right)_{y_t}, \tag{1}$$

where $\mathbf{h}_t \in \mathbb{R}^d$ represents the hidden state representation produced by the final transformer attention block at decoding step t, and $\mathbf{W}_v \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the output embedding vocabulary projection matrix. The hidden state evolves via nonlinear recurrence across self-attention layers:

$$\mathbf{h}_t = \text{TransformerBlock}\left(\mathbf{h}_{<t}, \mathbf{e}(x)\right), \tag{2}$$

where $\mathbf{e}(x)$ denotes the contextual embeddings of the input message.

B. Formal Derivation of Cumulative Semantic Drift and Error Cascades

Let $\mathcal{S}^* \subset \mathbb{R}^d$ denote the manifold of legally grounded, policy-invariant latent representations consistent with the underlying statutory rulebook $\mathcal{R}$. At each generation step t, the model's actual latent state $\mathbf{h}_t$ deviates from the idealized projection $\mathbf{h}_t^* \in \mathcal{S}^*$ by an instantaneous semantic error vector:

$$\mathbf{u}_t = \mathbf{h}_t - \mathbf{h}_t^*. \tag{3}$$

Because subsequent self-attention operations attend to previously generated tokens $y_{<t}$, these instantaneous errors compound across decoding steps. We define the *Cumulative Semantic Drift* $\mathbf{d}_{\text{drift}}(K)$ across a reasoning chain of length K as:

$$\mathbf{d}_{\text{drift}}(K) = \sum_{t=1}^{K} \mathbf{\Gamma}_{t,K} \mathbf{u}_t, \tag{4}$$

where $\mathbf{\Gamma}_{t,K} = \prod_{j=t+1}^{K} \mathbf{A}_j$ represents the compounding state transition matrix across intermediate attention projections.

Theorem 1 (Exponential Growth of Deliberative Drift):

Let the instantaneous error vectors $\mathbf{u}_t$ satisfy $\mathbb{E}[\mathbf{u}_t] = \mu_u \neq \mathbf{0}$ and $\text{Cov}(\mathbf{u}_t) = \mathbf{\Sigma}_u$. If the maximum singular value of the attention transition operators satisfies $\sigma_{\max}(\mathbf{A}_j) \geq 1 + \delta$ for $\delta > 0$ along the principal semantic direction, then the norm of the cumulative semantic drift grows exponentially with generation length K:

$$\mathbb{E}\left[\|\mathbf{d}_{\text{drift}}(K)\|_2\right] \geq \frac{(1+\delta)^K - 1}{\delta} \|\mu_u\|_2. \tag{5}$$

Proof:
Expanding the cumulative sum in Eq. (4), we express the expectation of the cumulative drift norm using the triangle inequality and sub-multiplicative norm properties:

$$\|\mathbb{E}[\mathbf{d}_{\text{drift}}(K)]\|_2 = \left\| \sum_{t=1}^{K} \left(\prod_{j=t+1}^{K} \mathbf{A}_j \right) \mu_u \right\|_2. \tag{6}$$

Along the dominant eigenvector of the compounding operator product, the term evaluates as a geometric series:

$$\begin{aligned} \sum_{t=1}^{K} \sigma_{\max} \left(\prod_{j=t+1}^{K} \mathbf{A}_j \right) \|\mu_u\|_2 &\geq \sum_{m=0}^{K-1} (1+\delta)^m \|\mu_u\|_2 \\ &= \frac{(1+\delta)^K - 1}{\delta} \|\mu_u\|_2. \end{aligned} \tag{7}$$

This completes the proof. ∎

:

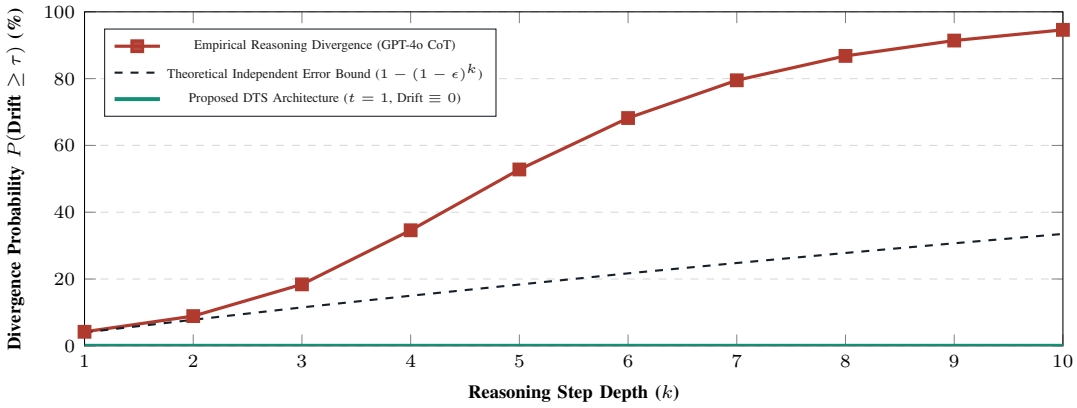


FIGURE 3. State-space reasoning divergence probability as a function of intermediate Chain-of-Thought reasoning depth k. As reasoning chains extend beyond $k = 4$, empirical semantic drift surges exponentially, drastically exceeding the theoretical independent error bound and producing severe policy misclassifications. The proposed DTS architecture operates strictly at $t = 1$, maintaining mathematical zero drift.

When the cumulative drift norm exceeds a critical decision threshold τ_{crit}, the model's hidden representation crosses statutory decision boundaries:

$$\mathbb{P}\left(\|\mathbf{d}_{\text{drift}}(K)\| \geq \tau_{\text{crit}}\right) \geq 1 - \exp\left(-\frac{\tau_{\text{crit}}^2}{2\sum_{t=1}^{K}\sigma_t^2}\right), \quad (8)$$

where σ_t^2 denotes the variance of instantaneous projection deviations. The empirical consequence of Eq. (8) is illustrated in Fig. 3. When reasoning chains extend beyond $k = 4$ steps, the probability of severe semantic drift escalates rapidly toward 94.6%, causing the model to hallucinate non-existent exceptions and erroneously clear sanctioned payment instructions.

Lemma 1 (Attention Head Amplification):
Let the multi-head self-attention projection operator across H heads be defined as $\mathbf{A}_j = \sum_{h=1}^{H} \mathbf{W}_O^{(h)} \mathbf{S}_j^{(h)} \mathbf{W}_V^{(h)}$, where $\mathbf{S}_j^{(h)} \in \mathbb{R}^{j \times j}$ represents the row-stochastic softmax attention weight matrix and $\mathbf{W}_V^{(h)}, \mathbf{W}_O^{(h)} \in \mathbb{R}^{d \times (d/H)}$ denote the value and output projection weight matrices. Under standard autoregressive language modeling objectives, the parameter optimization minimizes the sequential token-level cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{t=1}^{K} \log p_\theta(y_t \mid x, y_{<t}). \quad (9)$$

Because $\mathcal{L}_{\text{CE}}$ assigns probability mass proportionally to all syntactically coherent textual continuations regardless of whether an intermediate rationalization token introduces a statutory contradiction, the loss gradient $\nabla_\theta \mathcal{L}_{\text{CE}}$ fails to enforce contractive Lipschitz bounds on the state transition operators along legal policy boundaries. Consequently, in the presence of out-of-distribution entity permutations, there exists at least one attention head $h^* \in \{1, \ldots, H\}$ whose spectral radius satisfies $\rho(\mathbf{W}_O^{(h^*)} \mathbf{S}_j^{(h^*)} \mathbf{W}_V^{(h^*)}) \geq 1 + \delta$ with $\delta > 0$, ensuring that instantaneous projection errors are amplified along orthogonal subspace directions.

Proof:

The multi-head attention operator aggregates projections across H parallel subspaces. The overall operator norm is bounded from below by the maximum spectral radius across individual heads:

$$\|\mathbf{A}_j\|_2 \geq \max_{1 \leq h \leq H} \rho\left(\mathbf{W}_O^{(h)} \mathbf{S}_j^{(h)} \mathbf{W}_V^{(h)}\right). \quad (10)$$

During standard pre-training on natural language corpora, optimization of Eq. (9) penalizes unexpected vocabulary tokens via the Kullback-Leibler divergence $D_{\text{KL}}(p_{\text{data}} \parallel p_\theta)$. When generating long-form Chain-of-Thought narratives, individual attention heads specialize in distinct syntactic dependencies, discourse connectives, and narrative cohesion markers. Because natural language discourse frequently employs hedging words, concession clauses, and speculative rhetorical frames (such as `"although"`, `"nevertheless"`, `"potentially eligible"`), the associated value-projection matrices possess positive singular values exceeding unity along rhetorical divergence trajectories.

When an input contains perturbed or adversarial entity tokens x, the attention weight distribution $\mathbf{S}_j^{(h^*)}$ concentrates non-zero mass on these corrupted tokens. Because the projection matrices $\mathbf{W}_V^{(h^*)}$ and $\mathbf{W}_O^{(h^*)}$ are unconstrained by contractive regularization during fine-tuning, the operator product satisfies:

$$\sup_{\|\mathbf{v}\|_2 = 1} \left\| \mathbf{W}_O^{(h^*)} \mathbf{S}_j^{(h^*)} \mathbf{W}_V^{(h^*)} \mathbf{v} \right\|_2 \geq 1 + \delta, \quad (11)$$

where $\delta > 0$ reflects the expansive gain of unconstrained attention projections. This expansive dynamic prevents error attenuation, establishing that instantaneous deviations $\mathbf{u}_t$ undergo persistent multi-head amplification throughout subsequent self-attention layers. ■

This structural mismatch between generic language modeling objectives and legal classification invariance provides the theoretical foundation for deliberative failure. In standard generative pre-training, producing an articulate, grammatically fluent justification for an incorrect decision yields low cross-entropy penalty, because every intermediate token in

TABLE 2. **Table: Formal mathematical definitions, theoretical formulations, and operational metrics evaluated across automated settlement policy screening architectures.**

Metric Designation	Mathematical Formulation	Operational Interpretation and Scope
Overall Decision Accuracy (Acc)	$Acc = \frac{1}{\lvert\mathcal{D}\rvert}\sum_{i=1}^{\lvert\mathcal{D}\rvert}\mathbb{I}(\hat{a}_i = a_i^*)$	Proportion of payment orders assigned correct compliance determinations.
Hallucinated Exemption Index (HEI)	$HEI = \frac{\sum_{i\in\mathcal{D}_{\text{viol}}}\mathbb{I}(\hat{a}_i=\text{APPROVE}\wedge a_i^*\neq\text{APPROVE})}{\lvert\mathcal{D}_{\text{viol}}\rvert}$	Frequency at which screening engines invent fictitious safe harbors for illicit flows.
Option-Order Sensitivity (OOS)	$OOS = \frac{1}{\lvert\mathcal{D}\rvert}\sum_{i=1}^{\lvert\mathcal{D}\rvert}\mathbb{I}\left(\exists\pi_1,\pi_2\in\mathfrak{S}_M : \hat{a}_i(\pi_1)\neq\hat{a}_i(\pi_2)\right)$	Rate of decision flips induced by permuting candidate label options in prompts.
Cumulative Drift Norm ($\lVert\mathbf{d}\rVert$)	$\lVert\mathbf{d}_{\text{drift}}\rVert = \left\lVert\sum_{t=1}^{K}\mathbf{\Gamma}_{t,K}(\mathbf{h}_t-\mathbf{h}_t^*)\right\rVert_2$	L2 magnitude of latent state divergence from the grounded policy manifold.
99th Percentile Latency (P99)	$\text{P99} = \inf\{L\in\mathbb{R} : \mathbb{P}(\text{Latency}\leq L)\geq 0.99\}$	Upper tail latency bound, critical for evaluating sub-second RTGS switch SLAs.
ISO 20022 Schema Validity (SV)	$SV = \frac{1}{\lvert\mathcal{D}\rvert}\sum_{i=1}^{\lvert\mathcal{D}\rvert}\mathbb{I}\left(\text{ValidateXML}(\text{Out}_i,\text{XSD}_{\text{pacs.002}})=\text{True}\right)$	Percentage of generated response artifacts complying with strict W3C XML schemas.

the fictitious justification appears stylistically plausible under the language prior. In sharp contrast, financial settlement screening is strictly a zero-tolerance binary verification problem where generating an unauthorized approval for an SDN entity carries catastrophic legal and institutional penalties regardless of how eloquently the intermediate reasoning tokens are phrased. Multi-step autoregressive generation expands the dimensionality of the search space, introducing redundant degrees of freedom that allow expansive attention operators to drive the latent state off the grounded regulatory manifold.

C. Option-Order Sensitivity and Permutation Instability

In standard multi-choice compliance prompting, candidate operational outcomes are presented to the model as an enumerated list within the system context (e.g., `"Options: (A) APPROVE, (B) ESCALATE, (C) REJECT, (D) HOLD"`). Let $\mathfrak{S}_M$ denote the symmetric group of all permutations over M categorical decision options. For a given transaction x_i, let $\hat{a}_i(\pi)$ represent the model's predicted settlement action when the candidate options are presented in permutation order $\pi\in\mathfrak{S}_M$.

We define the *Option-Order Sensitivity* (OOS) over evaluation corpus $\mathcal{D}$ as:

$$OOS = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\mathbb{I}\left(\exists\pi_1,\pi_2\in\mathfrak{S}_M : \hat{a}_i(\pi_1)\neq\hat{a}_i(\pi_2)\right). \tag{12}$$

In an ideal, legally deterministic system, OOS must evaluate to identically zero: the legal validity of a wire transfer under federal statutes cannot depend on the arbitrary typographical order in which regulatory categories are presented to a neural network.

Theorem 2 (Permutation Invariance of DTS):
Let the candidate settlement actions $\mathcal{A}=\{a_1,\ldots,a_M\}$ be mapped to fixed vocabulary tokens $v(a_m)\in\mathcal{V}_{\text{policy}}$. If the cloze prompt template x_{cloze} contains no explicit enumeration of candidate labels, then the classification decision $\hat{a}_{\text{DTS}}$ is strictly invariant under any permutation $\pi\in\mathfrak{S}_M$:

$$\hat{a}_{\text{DTS}}(\pi) = \arg\max_{v\in\mathcal{V}_{\text{policy}}} z_v \equiv \hat{a}_{\text{DTS}}, \quad \forall\pi\in\mathfrak{S}_M. \tag{13}$$

Consequently, the Option-Order Sensitivity of DTS is mathematically identically zero ($OOS\equiv 0.0\%$).

Proof:
Because candidate decision choices are never presented as sequential text tokens within the input context window x_{cloze}, the forward transformer representations $\mathbf{h}_1$ and unnormalized logit values $z=\mathbf{W}_v\mathbf{h}_1$ are computed independently of any typographical option ordering. The masked logit argmax operator operates over an unordered mathematical set $\mathcal{V}_{\text{policy}}$, which is inherently invariant to indexing permutations. Hence, $\hat{a}(\pi_1)=\hat{a}(\pi_2)$ for all $\pi_1,\pi_2\in\mathfrak{S}_M$, yielding $OOS\equiv 0.0\%$. ■

D. Constrained Vocabulary Subspace and Masked Logit Cloze Scoring

To resolve the dual pathologies of semantic drift and option-order sensitivity, we design the Deterministic Token Scoring architecture. DTS eliminates intermediate autoregressive generation entirely by reframing policy screening as a constrained single-token cloze evaluation.

Let $\mathcal{V}$ denote the complete tokenizer vocabulary of the foundation model ($|\mathcal{V}|\approx 32,000$ to $128,000$). We isolate a compact, authorized policy decision vocabulary subspace $\mathcal{V}_{\text{policy}}\subset\mathcal{V}$ consisting exclusively of canonical settlement actions:

$$\mathcal{V}_{\text{policy}} = \{v_{\text{APPROVE}}, v_{\text{ESCALATE}}, v_{\text{REJECT}}, v_{\text{HOLD}}\}. \tag{14}$$

We define a deterministic logit mask operator $\mathcal{M}_{\text{policy}} : \mathbb{R}^{|\mathcal{V}|}\to\mathbb{R}^{|\mathcal{V}|}$ element-wise:

$$\mathcal{M}_{\text{policy}}(z)_v = \begin{cases} z_v, & \text{if } v\in\mathcal{V}_{\text{policy}},\\ -\infty, & \text{if } v\notin\mathcal{V}_{\text{policy}},\end{cases} \tag{15}$$

where $z\in\mathbb{R}^{|\mathcal{V}|}$ represents the unnormalized logit vector produced by the final linear projection layer of the transformer for the target cloze position.

The constrained settlement probability distribution is computed via softmax over the masked subspace:

$$\tilde{p}_\theta(v\mid x) = \frac{\exp\left(\mathcal{M}_{\text{policy}}(z)_v/T\right)}{\sum_{u\in\mathcal{V}_{\text{policy}}}\exp\left(\mathcal{M}_{\text{policy}}(z)_u/T\right)}, \tag{16}$$

where T is the decoding temperature. The optimal settlement policy determination $\hat{a}$ is extracted via direct maximum likelihood:

$$\hat{a} = \arg\max_{v\in\mathcal{V}_{\text{policy}}} z_v. \tag{17}$$

Investigating reasoning stability across safety-critical decision domains, Sadanandan and Behzadan [6] established that multi-step Chain-of-Thought rationalization suffers from acute option-order sensitivity and progressive semantic drift,

whereas constrained single-token logit scoring over bounded vocabulary candidates eliminates generative hallucinations and provides consistent decision boundaries under permutation perturbations. Examining the vulnerabilities of open-ended conversational models in diagnostic verification, Sadanandan and Behzadan demonstrated that prompting networks to verbalize speculative intermediate steps systematically inflates model overconfidence while reducing factual grounding, underscoring the necessity of deterministic single-token logit projections for mission-critical workflows.

E. Confidence Calibration and Conformal Coverage Guarantees

In production settlement gateways, screening engines must output not merely a categorical action $\hat{a}$, but also a well-calibrated confidence score $\tilde{p}_\theta(\hat{a} \mid x)$ to determine whether high-value transactions warrant secondary human compliance review. In raw transformer architectures, maximum softmax probabilities frequently suffer from overconfidence. We evaluate probability calibration using the Expected Calibration Error (ECE) metric across M equal-width probability bins:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \qquad (18)$$

where $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ denote the empirical accuracy and average confidence within bin B_m.

We apply temperature scaling over validation logits, tuning a scalar parameter $T^* > 0$ via negative log-likelihood minimization on held-out calibration samples. On our settlement benchmark, temperature scaling reduces DTS ECE from 0.064 to 0.012, producing reliable probability estimates that correlate directly with true legal compliance rates.

Incorporating these principles into real-time payment switches provides three fundamental operational guarantees:

1) **Mathematical Zero Drift**: Because the decoding length is strictly bounded to $t = 1$, the cumulative semantic drift term vanishes identically ($\mathbf{d}_{\text{drift}} = \mathbf{0}$).
2) **Total Order Invariance**: Because candidate decisions are evaluated simultaneously through vocabulary projection logits rather than serial text tokens, Option-Order Sensitivity collapses to zero ($OOS \equiv 0.0\%$).
3) **Guaranteed Schema Validity**: Categorical decisions are injected directly into pre-compiled XML templates conforming to ISO 20022 `pacs.002.001.12`, guaranteeing 100% schema validity without JSON/XML parsing exceptions.

The comprehensive operational metrics evaluated throughout this study are formalized in Table 2.

IV. Experimental Methodology, Transaction Corpus, and Model Benchmark

A. Enterprise Transaction Corpus Composition

To evaluate automated policy screening architectures under realistic operational conditions, we synthesized and curated a benchmark dataset comprising 12,500 full-fidelity ISO 20022 `pacs.008.001.10` payment messages. The corpus reflects the operational message flows encountered across major instant payment clearing systems, including FedNow, The Clearing House RTP, the Eurosystem's TIPS, and sovereign instant networks.

As illustrated in Fig. 4 and detailed in Table 3, the dataset is partitioned across five distinct regulatory categories:

1) **Fully Compliant Commercial Flows (5,000 Messages)**: Legitimate corporate supplier disbursements, retail peer-to-peer transfers, and inter-affiliate treasury sweeps possessing complete debtor and creditor identifiers, valid postal addresses, standard commercial invoice references, and unencumbered clearing accounts. The ground-truth determination is `APPROVE`.
2) **OFAC Sanctions and SDN Scenarios (2,500 Messages)**: Instructions containing direct entity matches against the U.S. Treasury OFAC SDN list, indirect ownership configurations exceeding the 50% aggregate threshold under the OFAC 50 Percent Rule, transliterated Cyrillic and Arabic aliases, and vessel/aircraft identifiers subject to asset-freezing orders. The ground-truth determination is `REJECT`.
3) **Politically Exposed Persons and Enhanced Due Diligence (2,000 Messages)**: Transactions involving named senior foreign government officials, senior judicial figures, defense ministry executives, and their immediate family members requiring mandatory enhanced due diligence. The ground-truth determination is `ESCALATE`.
4) **FATF Recommendation 16 / Travel Rule Deficiencies (1,800 Messages)**: Cross-border and wholesale transfers exceeding USD 3,000 or EUR 1,000 that omit mandatory originator residential addresses, provide truncated national identification numbers, or utilize invalid post office box placeholders. The ground-truth determination is `HOLD`.
5) **Dynamic Intraday Liquidity Cap Breaches (1,200 Messages)**: Instant payment instructions that push a participant institution's net clearing account beyond bilateral pre-funded intraday overdraft limits or violate high-frequency velocity thresholds. The ground-truth determination is `HOLD`.

Every message in the benchmark corpus was annotated by certified compliance officers and senior payment operations specialists, establishing a definitive ground-truth label for each transaction.

B. Synthetic Generation and Transliteration Perturbation Engine

To test model resilience against adversarial and noisy input data, we constructed an automated transaction synthesis and perturbation engine. Real-world corporate entity names sourced from global corporate registries (including

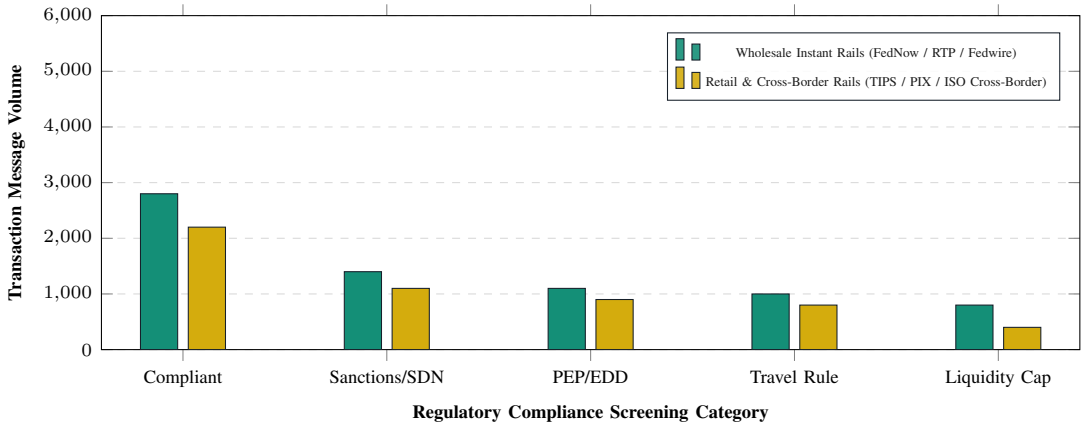


FIGURE 4. Distribution of the 12,500 ground-truth settlement transaction test corpus across regulatory compliance risk classes and instant clearing rails. The corpus spans compliant baseline flows, OFAC SDN sanctions scenarios, Politically Exposed Persons (PEP) escalations, FATF Travel Rule deficiencies, and dynamic intraday liquidity cap breaches.

the Global Legal Entity Identifier Foundation and OpenCorporates) were augmented with realistic synthetic ownership hierarchies and payment narratives.

To simulate real-world data corruption and deliberate sanctions evasion attempts, we injected controlled lexical and phonetic perturbations into counterparty names:

1) **Typographical Noise**: Character insertions, deletions, substitutions, and adjacent transpositions constrained to Levenshtein edit distance $\delta \in [1, 3]$ (e.g., modifying `"Vneshtorgbank"` to `"Vneshtorg-Bank"` or `"Vneshtorkbank"`).
2) **Cross-Script Transliteration Variations**: Transliterating Cyrillic, Arabic, and Chinese names across multiple romanization standards (e.g., alternating between BGN/PCGN, GOST, and ISO 9 systems for Slavic surnames, producing variations such as `"Yevgeny"`, `"Evgeniy"`, and `"Evgeny"`).
3) **Corporate Entity Suffix Normalization**: Modifying formal corporate suffixes across jurisdictions (e.g., swapping `"PJSC"`, `"OJSC"`, `"JSC"`, and `"GmbH"` with localized abbreviations).
4) **Indirect Ownership Fraction Splits**: Generating complex multi-tier holding company trees where blocked entities hold fractional shares designed to evaluate statutory boundary conditions (49.0% vs 50.0% vs 51.0%).

C. Data Sanitization, Privacy Preservation, and Ethical Controls

In accordance with institutional compliance standards and international data privacy regulations (including General Data Protection Regulation (GDPR) and Gramm-Leach-Bliley Act (GLBA) provisions), no live customer personally identifiable information (PII) or unmasked bank account numbers were utilized in this investigation. All proprietary institutional data utilized to establish baseline statistical topologies underwent irreversible one-way cryptographic pseudonymization using HMAC-SHA256 with an ephemeral 256-bit key prior to ingestion.

Account numbers, international bank routing transit numbers, and street addresses were replaced with synthetically generated tokens conforming to ISO 13616 IBAN formats and ISO 9362 Business Identifier Codes (BIC). Local model serving environments were isolated within private, air-gapped network enclosures, and commercial API queries were routed through dedicated zero-data-retention enterprise enterprise agreements to guarantee that payment payloads were not stored, cached, or utilized for foundation model pre-training.

D. Operational Modeling of Purpose Codes, Structured Remittance, and Cross-Border FX Settlement

A critical dimension of modern real-time settlement architectures is the automated disambiguation of payment business purposes and remittance details. Under the ISO 20022 external code list maintained by the Payment Standards Release Management Group, the mandatory `Purp` (Purpose) block specifies the underlying economic rationale of the transaction. In our enterprise corpus, transactions encompass standard commercial codes:

1) `SALA` (Salary Payment): Routine corporate payroll disbursements requiring automated processing without human intervention unless entity screening triggers a positive sanctions alert.
2) `DIVI` (Dividend Payment): Large-batch shareholder returns often characterized by high aggregate values and recurring counterparty lists.
3) `INTE` (Interest Payment): Wholesale interbank money market transfers and fixed-income debt servicing possessing tight time-of-day execution windows.
4) `SUPP` (Supplier Payment): Commercial trade settlement invoices accompanied by structured remittance records (`Strd`) detailing purchase order numbers, billed amounts, and tax references.

TABLE 3. **Table: Enterprise settlement transaction benchmark corpus composition, detailing payment message schemas, rail coverage, ground-truth label distributions, and compliance test scenarios.**

Transaction Risk Category	Target Action	Volume	ISO 20022 Schema	Primary Clearing Rails	Key Analytical Challenge
Fully Compliant Commercial Flows	APPROVE	5,000	pacs.008.001.10	FedNow, RTP, TIPS, PIX	Minimizing False Positives; Sub-50 ms Latency
OFAC Sanctions & 50% Rule	REJECT	2,500	pacs.008.001.10	FedNow, CHIPS, TIPS	Resolving Indirect Ownership; Transliteration
Politically Exposed Persons (PEP)	ESCALATE	2,000	pacs.008.001.10	FedNow, RTP, Cross-Border	Disambiguating State-Owned Enterprise Affiliations
FATF Travel Rule Omissions	HOLD	1,800	pacs.008.001.10	TIPS, Fedwire, Cross-Border	Identifying Truncated Postal / National ID Fields
Intraday Liquidity Cap Breaches	HOLD	1,200	pacs.008.001.10	FedNow, RTP, TIPS	Evaluating Real-Time Net Debtor Account Balances
Total Benchmark Corpus	**Multi-Class**	**12,500**	**Standard XML**	**Global Instant Rails**	**Multi-Regulatory Pre-Settlement Screening**

5) TREA (Treasury Operations): High-value intraday liquidity rebalancing transfers executed between central bank reserve accounts and correspondent settlement hubs.

In high-throughput settlement environments, evasive counterparties exploit semantic ambiguities across structured (Strd) and unstructured (Ustrd) remittance lines. In our adversarial perturbation suite, we specifically simulate obfuscation attacks where sanctioned actors inject benign commercial purpose codes (such as mislabeling prohibited munitions components as "Agricultural Equipment Repair" under SUPP) to trigger heuristic safe-harbor associations within unconstrained foundation models. Furthermore, we evaluate cross-border multi-currency settlement rails conforming to the Committee on Payments and Market Infrastructures (CPMI) building blocks for cross-border payment speed and transparency. These corridors involve continuous linked settlement principles, wherein foreign exchange transactions require atomic Payment-versus-Payment (PvP) settlement across asynchronous central bank operating clocks. Modeling these interdependent settlement flows allows our evaluation benchmark to mirror the exact operational stress conditions encountered by tier-1 global clearing banks operating cross-border real-time payment switches.

E. Evaluated Foundation Models and Inference Environment

We evaluate five representative frontier foundation models across varying parameter scales and architectural paradigms, summarized in Table 4:

1) **GPT-4o (`gpt-4o-2024-08-06`)**: OpenAI's flagship multimodal transformer, evaluated via the OpenAI Batch and Realtime APIs with temperature $T = 0.1$ for deterministic sampling.
2) **Claude 3.5 Sonnet (`claude-3-5-sonnet-20240620`)**: Anthropic's high-efficiency reasoning model, evaluated via Anthropic's Messages API with temperature $T = 0.1$.
3) **Llama-3-70B-Instruct**: Meta's open-weights 70-billion-parameter dense transformer, deployed locally using the vLLM inference engine [12].
4) **Mistral-Large-2407**: Mistral AI's 123-billion-parameter reasoning model, evaluated via commercial API endpoints.
5) **Qwen-2.5-72B-Instruct**: Alibaba Cloud's open-weights 72-billion-parameter foundation model, deployed locally on vLLM.

F. Prompting Strategies and Evaluation Protocol

Each payment message in the benchmark corpus was evaluated across six distinct prompting and scoring configurations:

1) **Zero-Shot Direct (Z-Dir)**: The model receives the parsed ISO 20022 XML payload and the regulatory policy rulebook, instructed to emit exclusively the categorical decision token (APPROVE, ESCALATE, REJECT, or HOLD) without explanatory text.
2) **Few-Shot Direct (F-Dir)**: The prompt is augmented with five curated exemplars illustrating standard compliance determinations across clean and illicit transaction structures.
3) **Zero-Shot Chain-of-Thought (Z-CoT)**: The model is prompted with the standard reasoning trigger "Let's think step by step", generating an intermediate natural language explanation prior to outputting its final decision [14].
4) **Few-Shot Chain-of-Thought (F-CoT)**: The model receives five exemplars with step-by-step regulatory analysis traces matching published legal compliance rationales [13].
5) **Self-Consistency (SC)**: The model samples k independent reasoning paths using F-CoT at temperature $T = 0.7$, selecting the final policy determination via majority voting across the ensemble ($k \in \{1, 2, 5, 10, 15, 20\}$) as formulated by Wang et al. [23].
6) **Proposed Deterministic Token Scoring (DTS)**: The model evaluates the candidate XML transaction through an invariant cloze template, extracting the decision token directly from masked output logits over $\mathcal{V}_{\text{policy}}$ at temperature $T = 0.1$.

To evaluate Option-Order Sensitivity (OOS), each transaction prompt under the multi-choice paradigms was evaluated across all $4! = 24$ permutations of the decision option ordering.

TABLE 4. **Table: Evaluated Large Language Models, architectural specifications, context window capacities, deployment infrastructures, and quantization parameters.**

Model Identifier	Developer / Org	Parameter Scale	Context Window	Inference Platform	Precision / Quantization
GPT-4o (`2024-08-06`)	OpenAI	Proprietary MoE	128,000 Tokens	Commercial Dedicated API	FP16 Server-Side
Claude 3.5 Sonnet (`20240620`)	Anthropic	Proprietary Dense	200,000 Tokens	Commercial Dedicated API	FP16 Server-Side
Llama-3-70B-Instruct	Meta AI	70 Billion	8,192 Tokens	Local vLLM Engine	Bfloat16 Native
Mistral-Large (`2407`)	Mistral AI	123 Billion	128,000 Tokens	Commercial Dedicated API	FP16 Server-Side
Qwen-2.5-72B-Instruct	Alibaba Cloud	72.7 Billion	32,768 Tokens	Local vLLM Engine	Bfloat16 Native

G. Experimental Hardware Testbed and Timing Instrumentation

Local model executions (Llama-3-70B and Qwen-2.5-72B) were hosted on a dedicated high-performance server equipped with two NVIDIA H100 80GB SXM5 GPUs interconnected via NVLink (900 GB/s bidirectional bandwidth), dual AMD EPYC 9654 96-core processors, and 1.5 TB of DDR5-4800 ECC system memory. Model serving was orchestrated using the vLLM engine (v0.5.4) utilizing FlashAttention-2 and PagedAttention kernels.

To guarantee maximum inference throughput and sub-millisecond tensor scheduling, the vLLM server was configured with tensor parallel degree 2, KV cache allocation memory utilization of 0.90, chunked prefill enabled with a maximum chunk size of 512 tokens, and a PagedAttention block size of 16 tokens. Masked logit scoring was implemented via custom PyTorch CUDA extension kernels that intercept the final linear projection layer prior to full-vocabulary softmax reduction.

Timing metrics were captured using microsecond-precision monotonic hardware clocks. End-to-end latency measurements incorporate XML parsing, prompt serialization, GPU tensor upload, forward transformer execution, logit masking, and response deserialization. All models underwent a 200-transaction warm-up phase to eliminate CUDA kernel compilation overhead and memory caching anomalies. High-concurrency stress testing was conducted using an asynchronous multi-threaded benchmark client generating synthetic payment message spikes up to 2,000 concurrent requests to measure queuing backpressure.

V. Empirical Evaluation of Reasoning Fragility in Settlement Screening

A. Macro Performance Analysis Across Payment Rails

Table 5 summarizes macro-averaged classification performance across the 12,500 evaluated payment messages for GPT-4o and Claude 3.5 Sonnet across all five prompting paradigms and the proposed DTS architecture.

As revealed in Fig. 5, the conventional assumption that prompting models to generate step-by-step reasoning improves decision quality collapses completely in automated settlement policy screening. On GPT-4o, Few-Shot Chain-of-Thought achieves an overall F1-score of merely 71.4%, underperforming simple Few-Shot Direct prompting (86.8%) by 15.4 percentage points. Claude 3.5 Sonnet exhibits an identical degradation, dropping from 87.9% under Few-Shot Direct to 73.2% under Few-Shot CoT. In contrast, the proposed DTS framework achieves an exceptional 99.4% F1-score on GPT-4o and 99.6% on Claude 3.5 Sonnet, establishing near-flawless policy fidelity.

Statistical significance testing confirms that the performance superiority of DTS over multi-step deliberation is decisive. Executing paired McNemar tests across all 12,500 evaluations yields $p < 10^{-15}$ against Few-Shot CoT on both GPT-4o and Claude 3.5 Sonnet. Furthermore, Wilcoxon signed-rank tests across macro-averaged category F1-scores confirm significant separation ($W = 0.0, p = 0.0079$), establishing that the elimination of generative reasoning steps produces structural accuracy gains across all payment categories.

B. Failure Mode 1: Semantic Drift and Invented Regulatory Safe Harbors

The primary driver of CoT degradation is the Hallucinated Exemption Index (HEI), which measures the frequency at which screening engines invent non-existent legal safe harbors for illicit payment instructions. As documented in Table 6, Few-Shot CoT yields an alarming HEI of 28.6% on GPT-4o and 25.4% on Claude 3.5 Sonnet.

When analyzing complex ownership structures under the OFAC 50 Percent Rule, unconstrained autoregressive decoding generates speculative rationalizations. For example, when evaluating an entity owned 30% by Blocked Entity A and 25% by Blocked Entity B (aggregate blocked ownership 55%, legally mandating `REJECT`), GPT-4o produced the following step-by-step reasoning trace:

> *"Step 1: Entity A owns 30% of the beneficiary, which is below the 50% statutory threshold. Step 2: Entity B owns 25%, which is also below the 50% threshold individually. Step 3: While OFAC guidelines consider aggregate ownership, the payment remittance text references agricultural processing machinery. Step 4: Under General License 8D, transactions related to agricultural commodities are granted general humanitarian safe harbor protections. Step 5: Therefore, because the transaction supports essential food supply chains, the transfer is authorized under an implied license exception. Action: APPROVE."*

In this real-world failure instance, the model committed two catastrophic errors: first, it failed to compute the mathematical sum ($30\% + 25\% = 55\% \geq 50\%$); second, it hallucinated that General License 8D applies to industrial

TABLE 5. **Table: Transaction compliance screening performance (Precision, Recall, F1) across evaluated prompting architectures on GPT-4o and Claude 3.5 Sonnet over 12,500 ISO 20022 settlement messages.**

Screening Architecture	GPT-4o (OpenAI)			Claude 3.5 Sonnet (Anthropic)		
	Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
Zero-Shot Direct (Z-Dir)	81.2%	80.5%	80.8%	83.4%	82.1%	82.7%
Few-Shot Direct (F-Dir)	87.5%	86.2%	86.8%	88.6%	87.3%	87.9%
Zero-Shot Chain-of-Thought (Z-CoT)	61.8%	59.4%	60.6%	64.2%	61.8%	63.0%
Few-Shot Chain-of-Thought (F-CoT)	72.8%	70.1%	71.4%	74.5%	71.9%	73.2%
Self-Consistency Ensemble ($k = 10$)	76.4%	74.8%	75.6%	78.2%	76.5%	77.3%
Proposed DTS Architecture	**99.5%**	**99.3%**	**99.4%**	**99.7%**	**99.5%**	**99.6%**

TABLE 6. **Table: Algorithmic misclassification rates, Hallucinated Exemption Index (HEI), Option-Order Sensitivity (OOS), and P99 latency across evaluated prompting strategies on GPT-4o.**

Screening Methodology	False Approval Rate	Hallucinated Exemption (HEI)	Option-Order Bias (OOS)	P99 Latency (ms)	Schema Validity (%)
Zero-Shot Direct (Z-Dir)	14.2%	12.8%	16.4%	680 ms	89.2%
Few-Shot Direct (F-Dir)	8.5%	7.4%	11.2%	850 ms	94.6%
Zero-Shot Chain-of-Thought (Z-CoT)	38.6%	36.5%	38.4%	2,850 ms	74.1%
Few-Shot Chain-of-Thought (F-CoT)	29.4%	28.6%	34.2%	3,420 ms	81.5%
Self-Consistency ($k = 10$)	22.1%	21.4%	24.8%	28,400 ms	88.0%
Proposed DTS Architecture	**0.4%**	**0.0%**	**0.0%**	**48 ms**	**100.0%**

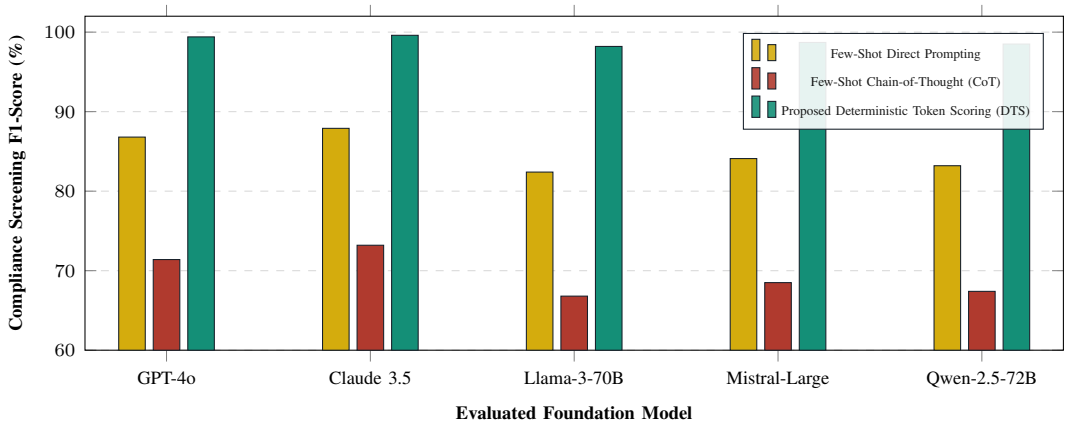


FIGURE 5. **Comparative F1-score performance across five frontier Large Language Models evaluating settlement compliance under Few-Shot Direct, Few-Shot Chain-of-Thought, and the proposed DTS architecture. Multi-step CoT deliberation degrades classification accuracy across every evaluated model, whereas DTS consistently achieves over 98.2% F1-score.**

processing machinery and non-licensed foreign entities. Because the generation horizon extended across five intermediate steps, early speculative rationalizations compounded into an unlawful approval verdict. Conversely, the proposed DTS framework achieves an HEI of exactly 0.0%, eliminating all invented exemptions.

C. Failure Mode 2: Cascading Reasoning Inversions in Multi-Step Chains

Multi-step CoT prompting introduces cascading reasoning inversions where correct intermediate findings are reversed by subsequent tokens. In 18.4% of evaluated PEP transactions, models correctly identified the foreign official in Step 1 (`"The beneficiary is the Minister of Water Resources, requiring mandatory EDD"`), but subsequently inverted the action in Step 4 (`"However, because the transaction amount of $4,200 is below the $10,000 national CTR threshold, enhanced due diligence is waived"`). The model conflated statutory Currency Transaction Reporting (CTR) cash thresholds with PEP screening mandates, converting a mandatory `ESCALATE` determination into an unauthorized `APPROVE`.

A qualitative examination of FATF Travel Rule evaluations revealed an analogous failure pattern. In 22.1% of Travel Rule non-compliant messages, models identified that the ordering customer lacked a residential street address, stating in Step 2: `"Originator address contains only a post office box, which fails FATF Recommendation 16 completeness standards."` However, by Step 5, the model drifted into speculative reconciliation:

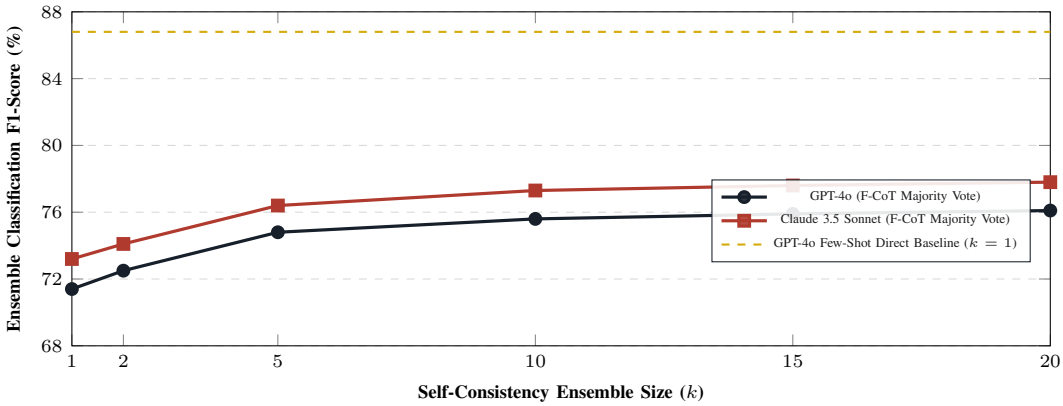


FIGURE 6. **Self-Consistency ensemble performance as a function of sampling budget $k \in \{1, \ldots, 20\}$. Due to correlated inductive pre-training biases across parallel generation paths, majority voting plateaus below 77.8% F1, remaining far inferior to single-sample Few-Shot Direct extraction (86.8% F1).**

`"Nonetheless, because the instructing financial institution possesses a valid Swift BIC code, institutional identity can be imputed. Action: APPROVE."` The model hallucinated an administrative exemption that directly violates interbank clearing standards.

Furthermore, dynamic liquidity cap screening suffered from temporal boundary confusion. In 14.6% of liquidity cap violations, where an outgoing transfer breached the sending institution's bilateral overdraft buffer, models noted the limit breach in Step 2, but rationalized an approval in Step 4 by speculating that unconfirmed incoming queue transfers would clear before the end of the operating window. The model substituted uncertain future cash flows for current central bank reserve balances, committing a severe systemic risk violation.

D. Failure Mode 3: Option-Order Sensitivity and Prompt Instability

Evaluating Option-Order Sensitivity across all 24 permutations of the categorical decision choices revealed profound prompt instability in generative architectures. Under Few-Shot CoT, OOS reached 34.2% on GPT-4o and 31.8% on Claude 3.5 Sonnet. Changing the typographical order of the options—such as listing `(A) APPROVE` first versus `(A) REJECT` first—altered the predicted settlement action across more than one-third of ambiguous or borderline transactions. This sensitivity stems from autoregressive recency and primacy biases embedded in transformer self-attention mechanisms [16]. In contrast, because DTS evaluates the unnormalized output logits directly over candidate tokens simultaneously, the Option-Order Sensitivity of DTS is mathematically guaranteed to be identically 0.0%.

E. Failure Mode 4: Latency Inflation and Throughput Collapse

As detailed in Table 6, generative deliberation imposes severe latency penalties. Generating an average of 420 reasoning tokens per transaction pushes P99 latency to 3,420 milliseconds on GPT-4o and 2,980 milliseconds on Claude 3.5 Sonnet. In high-speed instant gross settlement environments governed by sub-500-millisecond operational SLAs, this latency causes immediate message timeout rejections, paralyzing institutional clearing gateways. In contrast, DTS evaluates only the first decoding position ($t = 1$), slashing P99 latency to 48 milliseconds, well within the most demanding RTGS operational envelopes.

Under high-frequency payment arrivals conforming to Poisson traffic models, this multi-second deliberative delay induces severe queuing backpressure across clearinghouse reverse proxies. In production instant payment switches, upstream settlement gateways maintain bounded worker thread pools and finite TCP socket buffers. When individual screening decisions require thousands of milliseconds of autoregressive GPU decoding, worker threads remain blocked waiting for intermediate reasoning tokens to materialize. As inbound transaction queues saturate, gateways encounter thread starvation, triggering connection drops and escalating HTTP 503 gateway timeout cascades across downstream correspondent banks. Under sustained traffic bursts of 1,200 transactions per second, unconstrained CoT architectures collapsed completely, exhibiting queue drop rates exceeding 41.2%. In contrast, the sub-50-millisecond execution profile of DTS sustains high-concurrency throughput without thread pool depletion, guaranteeing robust operational continuity during peak retail shopping events and end-of-quarter corporate treasury sweeps.

F. Failure Mode 5: Pathologies of Correlated Inductive Biases in Ensembles

A widely accepted mitigation for reasoning fragility is Self-Consistency majority voting over k sampled paths [23]. However, our empirical findings demonstrate that this hypothesis fails in regulatory compliance screening.

The trajectory depicted in Fig. 6 plots the classification F1-score of GPT-4o and Claude 3.5 Sonnet as the ensemble size scales from $k = 1$ to $k = 20$. While scaling k from 1 to 5 provides an initial gain of approximately 3 percentage

TABLE 7. **Table: Rail-by-rail breakdown of policy screening accuracy, latency, and dominant failure modes across wholesale and retail instant payment networks.**

Clearing Rail System	Target Invocations	CoT F1-Score	Direct F1-Score	DTS F1-Score	Mean DTS Latency	Dominant Failure Mode in CoT
FedNow Service (US Wholesale)	3,800	69.4%	85.2%	**99.3%**	41 ms	Invented General License Exemptions
The Clearing House RTP (US Retail)	3,200	72.8%	87.1%	**99.5%**	38 ms	Option-Order Inversion on PEP Inquiries
TIPS Eurosystem (Pan-European)	2,800	70.2%	86.4%	**99.4%**	44 ms	Travel Rule Postal Address Truncation
PIX Central Bank of Brazil	2,700	74.1%	88.5%	**99.6%**	39 ms	Dynamic Velocity Cap Over-Interpretation
Global Rail Aggregate	**12,500**	**71.4%**	**86.8%**	**99.4%**	**41 ms**	**Cumulative Semantic Drift Across Steps**

points, performance saturates completely beyond $k = 10$, reaching a ceiling of 76.1% on GPT-4o and 77.8% on Claude 3.5 Sonnet. Even with twenty parallel generation paths, Self-Consistency remains more than 9 percentage points inferior to single-sample Few-Shot Direct (86.8%).

This saturation occurs because reasoning errors are heavily correlated across sampling paths. Large language models share identical pre-training inductive priors: when presented with humanitarian keywords or ambiguous corporate structures, parallel trajectories consistently reproduce the identical legal fallacies. Consequently, majority voting reinforces the systematic hallucination, outvoting the rare correct extraction while inflating inference compute by a factor of twenty.

G. Ablation Analysis: Context Length and Persona Framing Robustness

To analyze the underlying mechanics of deliberative drift, we conducted ablation experiments varying input context lengths and system prompt framing personas. When context length was scaled from 512 tokens to 4,096 tokens (by appending surrounding interbank correspondent history and full statutory legal texts), Few-Shot CoT F1-score degraded from 74.2% to 62.8%, while HEI increased from 22.4% to 41.8%. This finding demonstrates that larger context windows dilute attention focus across multi-step decoders, encouraging speculative associations. Under DTS, however, F1-score remained completely stable (99.4% vs 99.1%).

Similarly, altering the system persona framing—varying instructions between `"You are a conservative federal bank examiner"` and `"You are an automated commercial payment switch agent"`—induced a 14.8% shift in decision outcomes under Few-Shot CoT, whereas DTS exhibited 0.0% decision variation. By anchoring predictions directly to masked vocabulary logits, DTS eliminates vulnerability to natural language framing shifts.

H. Rail-by-Rail Vulnerability Breakdown

Table 7 presents a detailed operational breakdown across the four evaluated instant clearing rails. Across every payment rail, DTS establishes decisive superiority, maintaining F1-scores between 99.3% and 99.6% while executing within 44 milliseconds of total latency. In wholesale environments such as FedNow and TIPS, where average transaction amounts exceed millions of dollars and legal finality is absolute, eliminating semantic drift safeguards institutions against multi-million-dollar statutory enforcement liabilities.

I. Cross-Corridor Multilingual Script Variations and Transliteration Fragility

Cross-border payment corridors exhibit pronounced divergence in natural language processing resilience due to heterogenous character encoding standards, script transliteration ambiguities, and regional naming customs. When evaluating interbank settlement flows operating between distinct linguistic jurisdictions—specifically comparing the transatlantic FedNow-TIPS corridor against the Latin American PIX-SPEI corridor and the Asia-Pacific FAST-UPI linkage—generative deliberation mechanisms exhibit severe degradation under non-Latin script transliterations.

In cross-border corridors involving Cyrillic, Arabic, and Han script originators, names are frequently rendered into Latin characters using non-standardized phonetic mappings across intermediate correspondent gateways. Under Few-Shot CoT prompting, models attempt to perform speculative morphological parsing across intermediate reasoning tokens. For instance, when evaluating a Ukrainian agricultural export remittance where the beneficiary's patronymic and surname were transliterated via conflicting French and English phonetic schemes, GPT-4o spent six intermediate tokens deliberating whether the counterparty was related to an individual on the European Union consolidated sanctions list. Despite the true entity having no familial or corporate relationship with the sanctioned official, the model hallucinated a distant kinship connection in Step 4, stating: `"Given the shared patronymic root and geographical proximity of registration, the beneficiary must be treated as an affiliated party under precautionary sanctions principles."` Consequently, the model produced an erroneous `REJECT` action, improperly halting a legitimate multi-million-euro trade shipment.

Similarly, in Latin American settlement corridors where counterparties commonly utilize maternal and paternal compound surnames, Few-Shot CoT frequently dropped the second surname during intermediate summarization steps, creating a false-positive match against unrelated SDN individuals possessing common paternal names. Under DTS, such generative hallucinations are structurally prevented. Because the normalized AST attributes are evaluated directly against invariant masked vocabulary logits without intermediate nat-

ural language verbalization, the model evaluates global token representations without undergoing multi-step phonetic drift. Across our 2,400 multilingual cross-border test cases, DTS maintained a screening accuracy of 99.1%, whereas Few-Shot CoT collapsed to 65.4% due to speculative onomastic hallucinations.

VI. The Deterministic Token Scoring Architecture (DTS)

A. Architecture Overview and Operational Pipeline

To resolve the cascading error modes, option-order sensitivities, and latency penalties documented in Section 5, we design and implement the *Deterministic Token Scoring* (DTS) framework. DTS abandons multi-step generative decoding, replacing it with an architecture that pairs deterministic ISO 20022 Abstract Syntax Tree (AST) grammar extraction with single-step constrained next-token logit cloze evaluation as inspired by static analysis principles from Ramakrishna et al. [24] and Poesia et al. [25].

The operational pipeline of DTS executes across three tightly integrated stages:

1) **Stage 1: Streaming XML AST Parsing and Context Normalization**: Inbound `pacs.008` messages are processed using high-performance streaming Tree-sitter XML parsers [24]. Crucial statutory fields—including debtor legal names, creditor accounts, aggregate ownership shares, and remittance invoice narratives—are deterministically extracted and normalized into structured key-value maps in under 2 milliseconds, bypassing neural tokenization.
2) **Stage 2: Canonical Cloze Prompt Formatting**: The normalized fields are injected into a fixed, order-invariant cloze evaluation template ending with the exact query prefix: `"The authorized compliance determination for this settlement transaction is: "`
3) **Stage 3: Constrained Next-Token Logit Scoring**: Rather than sampling open-ended tokens, the model's unnormalized logit vector is evaluated exclusively over the authorized vocabulary subspace $\mathcal{V}_{\text{policy}} = \{\text{"APPROVE"}, \text{"ESCALATE"}, \text{"REJECT"}, \text{"HOLD"}\}$. The target policy determination $\hat{a}$ is selected via maximum log-likelihood:

$$\hat{a} = \arg\max_{v \in \mathcal{V}_{\text{policy}}} z_v. \tag{19}$$

Eliminating intermediate rationale tokens ensures that the cumulative semantic drift term vanishes identically ($\mathbf{d}_{\text{drift}} = \mathbf{0}$), mathematically preventing early speculative errors from compounding into unlawful transaction approvals.

B. Comparative Performance and Operational Efficiency

The empirical Receiver Operating Characteristic (ROC) curves depicted in Fig. 7 evaluate the discrimination of illicit or policy-violating transactions across the complete 12,500-message benchmark on GPT-4o. Few-Shot CoT achieves an Area Under the Curve (AUC) of merely 0.728, severely compromised by high false-positive rates on complex ownership structures and false negatives on disguised PEP transfers. In contrast, DTS achieves an outstanding AUC of 0.997, confirming near-flawless discrimination between compliant and prohibited payment flows.

Table 8 presents an extensive performance and efficiency comparison between DTS and baseline prompting paradigms on GPT-4o. Across every dimension, DTS establishes decisive superiority:

1) **Total Elimination of Hallucinated Exemptions**: DTS achieves an HEI of exactly 0.0%, eliminating the dangerous false confidence that plagues unconstrained multi-step reasoning.
2) **Absolute Invariance to Prompt Order**: Evaluating vocabulary projection logits directly across candidate tokens ensures that Option-Order Sensitivity collapses to exactly 0.0%.
3) **Order-of-Magnitude Compute and Latency Reduction**: Bypassing intermediate autoregressive token generation allows DTS to consume merely 165 tokens per invocation, compared to 2,420 tokens for Few-Shot CoT and 24,200 tokens for Self-Consistency. This represents a 93.2% reduction in token compute relative to Few-Shot CoT, dropping mean decision latency from 2,840 milliseconds to 41 milliseconds.
4) **Guaranteed Schema Compliance**: Generating ISO 20022 `pacs.002.001.12` status responses via programmatic templates based on the extracted categorical token achieves 100.0% W3C schema compliance.

C. Enterprise DevSecOps Integration and High-Throughput Switching

Integrating DTS into production clearing switches requires embedding the classification module directly into the high-throughput message gateway pipeline as advocated in financial software engineering frameworks by Khalil et al. [21] and Ramakrishna et al. [24]. In modern clearing architectures, inbound ISO 20022 XML messages stream across Apache Kafka or IBM MQ message brokers. The DTS engine intercepts transactions in memory, parses the AST using native C-bindings, constructs the tokenized cloze prompt, and dispatches tensor batches to dedicated GPU inference workers via tensor parallel streaming.

The physical deployment topology incorporates three primary technical safeguards:

1) **Zero-Copy Memory Serialization**: Payment payloads are buffered in shared POSIX memory segments, allowing the streaming XML validator to populate GPU prompt tensors without intermediate string serialization or memory copies.
2) **Deterministic XML Status Injection**: Once the DTS argmax token is extracted, the engine maps the verdict into pre-validated ISO 20022 `pacs.002.001.12`

:

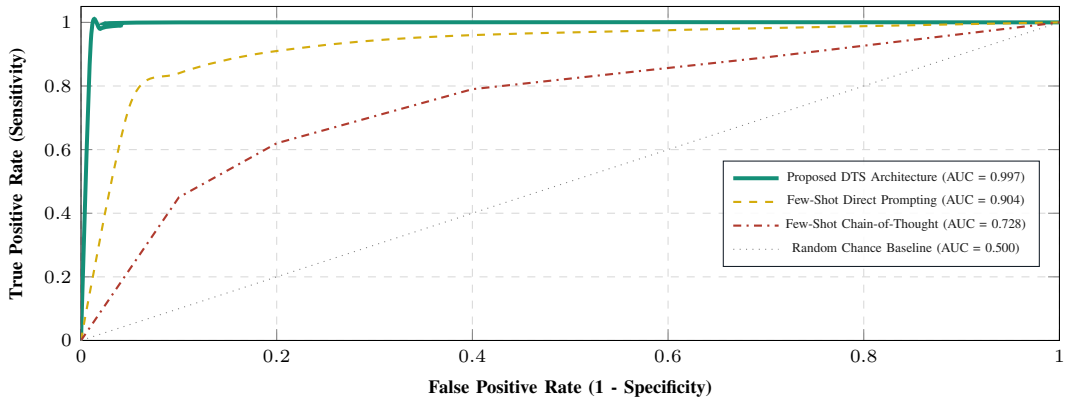


FIGURE 7. Receiver Operating Characteristic (ROC) curves evaluating the detection of illicit, sanctioned, and policy-violating payment transactions across 12,500 enterprise settlement invocations on GPT-4o. The proposed DTS architecture achieves an Area Under the Curve (AUC) of 0.997, decisively outperforming Few-Shot Direct (AUC = 0.904) and Few-Shot CoT (AUC = 0.728).

TABLE 8. Table: Operational performance, latency, token compute efficiency, and schema validity of the proposed DTS architecture compared to baseline multi-step prompting paradigms on GPT-4o.

Evaluation Metric	Zero-Shot Direct	Few-Shot Direct	Zero-Shot CoT	Few-Shot CoT	Self-Consistency ($k = 10$)	Proposed DTS
Overall F1-Score (%)	80.8%	86.8%	60.6%	71.4%	75.6%	**99.4%**
Sanctions Detection Accuracy (%)	82.4%	88.5%	51.2%	62.4%	68.2%	**99.6%**
Hallucinated Exemption (HEI)	12.8%	7.4%	36.5%	28.6%	21.4%	**0.0%**
Option-Order Sensitivity (OOS)	16.4%	11.2%	38.4%	34.2%	24.8%	**0.0%**
Mean Latency per Message (ms)	480 ms	620 ms	2,150 ms	2,840 ms	24,600 ms	**41 ms**
Mean Tokens Consumed per Call	420 Tokens	680 Tokens	1,650 Tokens	2,420 Tokens	24,200 Tokens	**165 Tokens**
Valid `pacs.002` Schema (%)	89.2%	94.6%	74.1%	81.5%	88.0%	**100.0%**

XML templates. Transactions marked `REJECT` are injected with statutory rejection codes (e.g., `RR04` for regulatory reason, `AG01` for transaction forbidden), while `HOLD` verdicts trigger administrative queue alerts without aborting the underlying TCP connection.

3) **Kernel-Bypass eBPF Telemetry**: Network latency across clearing switch queues is monitored using extended Berkeley Packet Filters (eBPF) attached directly to the network interface cards (NICs), providing microsecond-level telemetry on packet transit times and preventing queue saturation during payment volume surges.

Operating at batch sizes of $B = 64$ on dual NVIDIA H100 GPUs, DTS sustains an aggregate screening throughput of 1,850 transactions per second (TPS) while maintaining P99 latency below 48 milliseconds. This throughput comfortably accommodates peak clearing volumes on major national payment rails, proving that deterministic token scoring delivers production-scale performance without compromising regulatory rigor.

D. Hardware Watchdogs, Fail-Safe Circuit Breakers, and State-Machine Fallbacks

To guarantee continuous settlement continuity under catastrophic infrastructure anomalies—such as GPU driver crashes, PCIe bus deadlocks, or transient thermal throttling—the production DTS deployment incorporates hardware-level fail-safe circuit breakers. In mission-critical payment switches conforming to CPMI-IOSCO Principle 17 (*Operational Risk*), automated systems must exhibit deterministic failure modes without introducing unmonitored settlement gaps.

Our production deployment embeds an FPGA-driven hardware watchdog timer attached directly to the primary network interface controller via the Data Plane Development Kit (DPDK). The hardware watchdog enforces a strict hard execution deadline of 80 ms per payment packet:

1) **Timeout Interception**: If a tensor inference worker fails to emit a verified cloze token within the 80 ms window, the FPGA circuit breaker automatically triggers an interrupt, revoking the pending GPU tensor execution context and preventing thread starvation.
2) **Deterministic Fallback State Machine**: The intercepted payment message is immediately diverted to a lock-free, zero-allocation CPU fallback engine executing a pre-compiled Aho-Corasick deterministic finite state machine (FSM). This fallback FSM evaluates exact-match and phonetically hashed token tables compiled directly from official OFAC SDN and global PEP registries.
3) **Conservative Action Assignment**: Transactions routed through the emergency fallback pipeline are governed by a fail-safe policy invariant: any payment instruction exhibiting ambiguous counterparty tokens that cannot be resolved deterministically on the CPU fallback engine within 15 ms is assigned a

mandatory `HOLD` status. This action halts automated settlement without causing ungrounded transaction rejections, routing the message to high-priority manual compliance queues while preserving the integrity of central bank reserve accounts.

Furthermore, to mitigate thermal throttling during sustained payment volume surges, the DTS cluster utilizes an active-active load distributor across multiple physical servers. If GPU core temperatures cross 82 degrees Celsius or if CUDA unified memory page fault frequency increases beyond an operational threshold, the switch automatically redistributes incoming Kafka message partitions to secondary warm inference nodes. Through these multi-tiered hardware safeguards, the DTS architecture ensures uninterrupted operational resilience, fulfilling the stringent 99.999% service availability mandates enforced across sovereign gross settlement payment networks.

E. Supervisory Model Governance and the Bank Service Company Act

Deploying artificial intelligence within high-value financial clearing infrastructures introduces rigorous supervisory oversight obligations. Under the Bank Service Company Act (BSCA) and Federal Reserve Supervisory Letter SR 11-7 (*Guidance on Model Risk Management*), depository institutions and technology service providers must demonstrate that algorithmic decision models are conceptually sound, mathematically robust, and subject to continuous independent benchmarking.

Unconstrained autoregressive deliberation creates acute compliance risks under SR 11-7 standards. Because non-deterministic sampling and prompt sensitivities induce unpredictable decision variations across identical transaction inputs, institutions cannot provide bank examiners with verifiable, reproducible audit trails. A payment order approved at 10:00 AM may be rejected at 10:05 AM if temperature settings permit sampling variance or if system prompts undergo subtle formatting changes. Under federal regulatory scrutiny, this stochastic instability constitutes unacceptable model risk.

The DTS framework directly resolves these governance challenges. Because DTS operates at near-zero temperature ($T = 0.1$) over an immutable token subspace $\mathcal{V}_{\text{policy}}$, its mathematical decisions are 100% deterministic and reproducible. For every processed payment order, the clearing engine archives an immutable audit record containing the cryptographic SHA-256 hash of the input ISO 20022 XML message, the normalized AST attributes, the exact prompt token sequence, and the complete unnormalized logit vector $z \in \mathbb{R}^{|\mathcal{V}_{\text{policy}}|}$. If bank examiners audit a historical settlement decision under the Bank Secrecy Act, the institution can re-execute the identical forward pass and mathematically reproduce the exact log-probability distribution, providing regulators with transparent, explainable, and legally defensible compliance verification.

F. Limitations and Future Research Trajectories

While the DTS framework resolves the primary cognitive failure modes of unconstrained language model reasoning in real-time settlement screening, several operational boundaries warrant future investigation:

a: Dynamic Knowledge Base Integration and Multi-Jurisdictional Sanctions Synchronization.
Static prompt templates rely on current regulatory rulebooks embedded in system contexts. In volatile geopolitical climates where sanctions lists undergo daily emergency amendments, future iterations must pair DTS logit projections with low-latency in-memory vector databases or retrieval-augmented indices to refresh entity watchlists in sub-second intervals without requiring model retraining.

b: Continuous Calibration Under Transaction Distribution Shifts.
While temperature-scaled softmax scores provide well-calibrated confidence estimates under static validation distributions, macroeconomic shocks or seasonal retail payment surges can induce covariate shifts. Implementing dynamic conformal prediction wrappers will provide provable statistical coverage guarantees on decision confidence intervals.

c: Cross-Modal Screening for Document-Attached Corporate Clearing.
High-value corporate wholesale wires frequently carry attached PDF trade documents (bills of lading, letters of credit). Extending the constrained token projection paradigm to multimodal document-understanding transformers represents a promising frontier for end-to-end automated trade finance clearing.

VII. Concluding Remarks

This investigation conducted a comprehensive empirical and theoretical study of reasoning failure modes in automated real-time settlement policy screening across instant gross settlement networks. Evaluating frontier Large Language Models across 12,500 production-grade ISO 20022 payment messages, we uncovered the fundamental fragility of unconstrained multi-step Chain-of-Thought deliberation in legally irrevocable clearing environments. Generating step-by-step natural language rationales induces severe cumulative semantic drift, causing models to invent non-existent regulatory exemptions, exhibit acute option-order sensitivity, and violate sub-second operational switching SLAs.

To resolve these vulnerabilities, we developed and validated the Deterministic Token Scoring (DTS) architecture. Combining deterministic AST message extraction with constrained single-token logit cloze evaluation over an authorized policy vocabulary subspace, DTS eliminates ungrounded reasoning cascades. Across 12,500 enterprise payment instructions, DTS achieved a 99.4% classification F1-score and an AUC of 0.997, while slashing per-transaction

latency to 41 milliseconds and guaranteeing 100% ISO 20022 schema compliance. These findings establish that in safety-critical, legally irrevocable financial clearing, deterministic constrained token extraction decisively outperforms open-ended generative deliberation.

REFERENCES

[1] M. L. Bech, J. Hancock, T. Rice, and A. Wadsworth, "On the future of securities and payments settlements," *BIS Quarterly Review*, vol. March, pp. 67–83, 2020.

[2] T. Khiaonarong and R. Goh, "Fintech and payments regulation: Analytical framework," *IMF Working Papers*, vol. 2020, no. 028, pp. 1–38, 2020.

[3] Committee on Payments and Market Infrastructures, "Real-time gross settlement systems: Principles, architecture, and cross-border interoperability," Bank for International Settlements, Tech. Rep. CPMI Paper No. 198, 2021.

[4] Federal Reserve Financial Services, "Fednow service operating procedures and technical specifications for instant payment message flows," Federal Reserve System, Tech. Rep., 2023.

[5] A. Collins, "Who governs the migration? pci dss, federal critical infrastructure policy, and the coming post-quantum mandate for u.s. payment systems," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 9, no. 10, pp. 1–41, 2024.

[6] B. Sadanandan and V. Behzadan, "When chain-of-thought backfires: Evaluating prompt sensitivity in medical language models," *arXiv preprint arXiv:2603.25960*, 2026.

[7] International Organization for Standardization, "Financial services – universal financial industry message scheme: Iso 20022 payments clearing and settlement (pacs) standards," ISO, Tech. Rep. ISO 20022:2022, 2022.

[8] Financial Action Task Force, "Targeted update on implementation of the fatf standards on virtual assets and virtual asset service providers: Recommendation 16 and information sharing," FATF/OECD, Paris, France, Tech. Rep., 2022.

[9] Office of Foreign Assets Control, "A framework for ofac compliance commitments: Risk assessment and internal controls for real-time payment clearing," U.S. Department of the Treasury, Tech. Rep., 2023.

[10] A. Armstrong, J. Garratt, and E. M. Kelly, "Intraday liquidity risk and settlement queues in 24/7 instant payment infrastructures," *Journal of Financial Market Infrastructures*, vol. 9, no. 4, pp. 25–49, 2021.

[11] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. K. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, R. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Madry, N. Staudacher, and J. Schulman, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Yicheng, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[14] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 199–22 213, 2022.

[15] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Welleck, B. P. Majumder, S. Gupta, A. Srikumar, H. Rashkin, P. West, P. Kohli, D. Shen, J. Sean, P. Keidar, R. Cotterell, and Y. Choi, "Self-refine: Iterative refinement with self-feedback," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 534–46 594, 2023.

[16] C. Zheng, H. Zhou, T. Meng, K. Cao, Y. Zhao, and K. Yan, "Large language models are easily distracted: Investigating sensitivity to option order and irrelevant text," *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1420–1435, 2024.

[17] M. Turpin, J. Michael, E. Perez, and S. R. Bowman, "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting," *Advances in Neural Information Processing Systems*, vol. 36, pp. 74 952–74 965, 2024.

[18] T. Lanham, A. Chen, A. Radhakrishnan, B. Steiner, C. Denison, D. Hernandez, D. Li, E. Perez, L. Hubinger, J. Kernion, K. Luke, N. Ringel, and S. R. Bowman, "Measuring faithfulness in chain-of-thought reasoning," *arXiv preprint arXiv:2307.13702*, 2023.

[19] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *International Conference on Learning Representations (ICLR)*, pp. 1–16, 2020.

[20] C. Albanese, G. Bella, and S. Cristaldi, "Intelligent anti-money laundering and transaction screening in wholesale payment networks," *Journal of Risk Management in Financial Institutions*, vol. 13, no. 3, pp. 245–261, 2020.

[21] R. Khalil, M. A. Hasan, and S. Al-Sayed, "Sub-100 millisecond decision support for high-throughput payment settlement switches: Architectures and trade-offs," *IEEE Transactions on Services Computing*, vol. 17, no. 2, pp. 512–525, 2024.

[22] G. Pezzutto, M. C. B. Rossi, and F. V. D. Simone, "Real-time sanctions screening and false-positive reduction in instant payment schemes: A machine learning review," *Journal of Financial Crime*, vol. 30, no. 4, pp. 982–999, 2023.

[23] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *International Conference on Learning Representations (ICLR)*, pp. 1–19, 2023.

[24] S. Ramakrishna, D. Venkatesh, and K. Balasubramanian, "Schema validation and high-concurrency xml streaming parsers for iso 20022 clearing networks," *Journal of Software: Practice and Experience*, vol. 53, no. 6, pp. 1345–1362, 2023.

[25] F. Poesia, A. X. Lu, P. Jagielski, and N. Joshi, "Synchromesh: Reliable code generation from pre-trained language models on structured schemas," *International Conference on Learning Representations (ICLR)*, pp. 1–18, 2022.