Why Large Language Models Fail at BNPL Logic — and How a Deterministic Co-Processor Can Fix It

Sanjay Mittal and Pranav Tonpe
Predictika.ai

Executive Summary – Overview

We conducted a structured evaluation of Buy Now Pay Later (BNPL) reasoning using representative Klarna policy scenarios spanning refunds, eligibility, cross-border usage, disputes, fees, defaults, and multi-product interactions. Each test case had a clearly defined expected outcome derived directly from Klarna contractual terms.

We evaluated three common deployment patterns: (1) API LLM with no rules provided; (2) Web UI LLM with rules injected via system prompts; and (3) API LLM with rules injected via system prompts. Across 31 consolidated test cases, accuracy varied dramatically: approximately 42% for API without rules, 20% for Web UI with rules, and 77% for API with rules. Even in the best configuration, nearly one in four answers was incorrect or indeterminate.

Failures impacted core BNPL operations including refund timing, installment cancellation, eligibility after default, merchant category restrictions, fee caps, arbitration windows, governing law, cross-border availability, and product-specific overrides.

What Actually Goes Wrong

LLMs frequently hedge instead of deciding. When scenarios require applying multiple rules, models default to responses such as “not enough information,” even when documents explicitly specify outcomes.

LLMs also exhibit over-literal interpretation. If a policy does not state a conclusion verbatim, models often refuse to infer logical consequences

(c) Predictika.ai 2026

that are operationally obvious.

Many BNPL scenarios require combining eligibility criteria, product terms, fee schedules, and jurisdictional clauses. Models typically understand individual fragments but fail to synthesize them into a single coherent decision.

Finally, prompt fragility is severe. Web UI deployments showed accuracy as low as 20% despite identical rules being injected, demonstrating that system prompts are not a reliable governance mechanism.

Why This Matters

BNPL systems operate at the intersection of consumer finance, regulatory compliance, contractual obligations, risk management, and customer experience. Approximate reasoning is unacceptable. Errors directly translate into regulatory exposure, financial losses, customer disputes, reputational damage, and legal risk.

Even when provided exact rules, LLMs fail to apply them consistently. This is not a tuning problem—it is an architectural limitation.

Estimated Annual BNPL Query Volume and Business Impact of LLM Errors

To understand the real-world implications of these accuracy gaps, it is important to estimate the scale at which BNPL-related decisions and customer queries occur.

In 2024, global BNPL transaction volume is estimated to exceed $500–600 billion, spanning both online and in-store purchases. Industry data indicates roughly 8–10 billion BNPL transactions annually worldwide, driven by more than 400 million active BNPL users and several million participating merchants. Each transaction typically generates multiple

(c) Predictika.ai 2026

logic-driven interactions including eligibility checks, payment-plan calculations, refund inquiries, dispute handling, installment changes, and customer support questions.

Conservatively assuming 3–5 policy-driven queries per BNPL transaction, this implies approximately 25–40 billion BNPL logic queries per year across digital channels, merchant integrations, mobile apps, and call centers.

Applying the experimentally observed error rates from this study:

• API LLM with rules (~77% accuracy) implies 6–9 billion incorrect or indeterminate decisions annually.
• API LLM without rules (~42% accuracy) implies 14–23 billion failures.
• Web UI LLM with rules (~20% accuracy) implies 20–32 billion failures.

These failures directly affect financial outcomes including incorrect eligibility determinations, refund and installment errors, misapplied fees or arbitration rules, and ambiguous responses that escalate support volume.

Even if only a small fraction of these errors result in disputes or manual intervention, the aggregate impact easily reaches hundreds of millions to billions of dollars annually in operational expense, compliance exposure, and revenue leakage. At tens of billions of decisions per year, even single-digit percentage improvements translate into massive financial and customer-experience gains.

Deterministic execution of BNPL policies is therefore not an optimization—it is a structural requirement.

The Deterministic Co-Processor Model

(c) Predictika.ai 2026

Predictika introduces a deterministic symbolic logic engine that operates as a co-processor alongside any LLM. In this architecture, the LLM handles natural language understanding, conversational flow, and intent extraction, while Predictika performs rule execution, constraint validation, eligibility determination, state transitions, and compliance checks.

The LLM becomes the language interface. Predictika becomes the reasoning engine.

How Predictika Can Fix BNPL Failures

Predictika encodes policies as executable logic rather than prompts, guaranteeing rule execution. It natively supports multi-rule composition through inheritance, overrides, scoped constraints, and temporal conditions.

Instead of ambiguous responses, Predictika produces deterministic Pass, Flag, or Block outcomes. Every decision includes applied rules, evaluated conditions, and final results, creating a complete audit trail.

Architectural Impact

In production, the flow is straightforward: a user asks a question; the LLM extracts intent and parameters; Predictika evaluates deterministically; Predictika returns structured outcomes; and the LLM renders the response in natural language. This eliminates hallucinated logic and enforces contractual correctness across channels.

Strategic Implications

LLMs alone are unsuitable for enterprise systems involving rules, rights, eligibility, money, or regulation. BNPL is only one example. The same failure modes appear in banking compliance, insurance underwriting, CPQ, claims processing, corporate governance, and government services.

(c) Predictika.ai 2026

Predictika enables a new standard: LLM plus deterministic logic equals enterprise-grade AI.

Closing

The Klarna BNPL experiments show that even with official rules injected, modern LLMs fail in roughly 25–80% of realistic financial scenarios depending on deployment mode. This gap cannot be closed with better prompts. It requires a co-processor.

Predictika provides that missing layer, delivering deterministic reasoning, compliance guarantees, and auditability for AI agents operating in high-stakes domains. This is how enterprise AI must be built.

Appendix – LLM Performance Summary by Scenario

Scenario A (API + Klarna Rules): 77% accuracy, driven by multi-rule synthesis and over-literal interpretation failures.
Scenario B (Web UI + Klarna Rules): 20% accuracy, primarily due to prompt dilution and rule ignoring.
Scenario C (API without Rules): 42% accuracy, dominated by hedging and missing contractual logic.

Selected Publications & Suggested Reading

Logic Engines to Guide LLMs for Accurate and Reliable Reasoning — Predictika White Papers (2025-26)
https://predictika.com/assets/doc/Guiding_AI_agents_to_reason_correctly.html

https://predictika.com/assets/doc/Accuracy_is_not_consistency.html
https://predictika.com/assets/doc/Logic%20_engine_to_guide_LLM.html
https://predictika.com/assets/doc/Where_is_AGI.html

(c) Predictika.ai 2026

EU Travel Compensation Bot

U.S. Patents in Conversational AI, Symbolic Reasoning, and LLM Governance (granted 2023–2024)

(c) Predictika.ai 2026