comviva
A TECH MAHINDRA COMPANY

Vritti
October 2025

TABLE OF CONTENTS

1. AI's renaissance moment - by Rajesh

How AI is transforming transaction into an "Experience"

2. AI Use case - by Sandhya

Smart Budgeting

3. AI in Engineering - by Swagata

Finding the brilliant middle

4. FinTech's Competitive Edge - by Gopal

AI as the Differentiator

5. The AI revolution is here - by Ved

What is your take?

6. Grafana with AI - by Ajit

Turning Monitoring into Business Intelligence

7. AI Agents in Java - by Mohit

Automating Non-Functional Requirements for Secure, Scalable Apps

8. Recent Awards & Accolades

Preface

Artificial Intelligence is no longer a distant possibility — it is here, transforming how we design, build, and experience technology. The pace of change is extraordinary, and at Comviva, we see this moment not as a disruption to react to, but as an opportunity to lead with purpose.

Across industries, AI is moving beyond efficiency and automation. It's enabling systems that can sense, learn, and adapt — turning transactions into experiences and data into decisions. For us, this represents a natural extension of our innovation journey. Every product we build, every solution we enhance, now carries an intelligent layer designed to make it smarter, faster, and more human in how it serves our customers.

This edition of Vritti brings together that collective vision. Through the lens of our experts, you'll see how AI is reshaping every facet of our work — from predictive budgeting and intelligent engineering to agentic fintech systems and self-optimizing platforms. These articles are not theoretical explorations; they reflect real innovation happening within Comviva, where technology and creativity meet to solve real-world problems.

Our focus remains clear — to make AI for all a reality across our product ecosystem. Whether it's enhancing customer engagement, improving operational efficiency, or creating new business models, AI is helping us move faster toward a world where every interaction is more intuitive and every outcome more intelligent.

I invite you to explore this edition and discover how Comviva is shaping the future — responsibly, inclusively, and with an unwavering commitment to innovation. The ideas shared here are not just about what AI can do, but about how we can use it to create technology that truly understands, empowers, and elevates human experience.

Enjoy the read — and join us in exploring how intelligence is becoming the new language of technology.

Srinivas Nidugondi

Chief Operating Officer, Digital Financial Solutions

01 AI's renaissance moment: How AI is transforming transaction into an "Experience" *by Rajesh*

With over 25 years of experience in Payments, digital channels and Transaction Banking, Rajesh has a proven track record of successfully managing digital business and driving transformation. He holds a deep understanding of the Retail as well as corporate digital channels, Transaction Banking and the evolving digital payments landscape. His expertise span across various domains, including retail banking, corporate banking, and trade finance. He is passionate about delivering exceptional digital experiences to customers and driving business growth through innovative solutions. Rajesh possesses strong leadership and interpersonal skills, enabling him to effectively collaborate with cross-functional teams and drive successful project outcomes.

Throughout history, significant technological epochs have followed a discernible pattern of innovation, transforming established systems and shaping the trajectory of human advancement. Gutenberg's press in the 15th century dissolved barriers around knowledge, redistributing power, authority and expanding access to knowledge throughout Europe and later to the rest of the world, changing the course of history. In more recent times, the internet stitched together the world's cognitive surplus, letting collaboration and information flow at unprecedented speed and at a global scale. Now, artificial intelligence arrives not as just another scientific advancement, but as a phenomenon threatening an equally historic reordering and perhaps ushering an "Industrial Revolution 2.0".

The shift from incremental upgrade to civilizational inflection

Historically, true breakthroughs are revealed less by what they do and more by what they make possible. Gutenberg's metal type didn't just create cheap books; it enabled religious reforms, democratization of knowledge leading to rise of science & learning, all within a few generations. The internet started as a humble research network and evolved into the backbone for commerce, industry, social life and the exchange of creativity. The core lesson: progress ceases being measured in speed or cost and instead begins rewriting foundational rules of 'who decides', 'who participates', 'who benefits'.

AI today, especially in its language model incarnation, already is transforming business and finance, amplifying productivity, automating tasks and creating new products with speed once unimaginable. Yet, as the Financial Times recently reports, leading companies like Elon Musk's xAI, Meta and Google are racing to develop "world models" – AI systems capable of understanding and navigating real – world physical environments. These next-generation systems train on video and robotic data to move beyond text-based understanding, marking a fundamental shift from digital reasoning to embodied intelligence.

World models: Beyond knowledge work to physical intelligence

Lately large language models (LLMs) move from curiosity to every layer of modern enterprise, with each new release showcasing new capabilities in text generation, reasoning and coding so rapid they almost seemed limitless. But as research focuses turn toward the maturity of systems like GPT-4, Claude and Gemini, the mood is shifting: notable progress in core reasoning or conceptual innovation has begun to slow and much of today's work refines user experience, safety and domain-specific reliability rather than upending previous boundaries. This plateau is not for lack of effort but architectural bottlenecks and exhausted high-quality data explain much of the deceleration, while transformer-based models excel at pattern recognition but stumble over real compositional abstraction or context-aware logic.

This juncture has sparked a new wave of ambition across the AI landscape. Rather than pursuing scale alone, researchers and businesses are now investing in architectures and datasets that reflect the physical world's complexity. "World models" are at the heart of this shift . They are AI systems trained to build dynamic internal maps of environments, sense, act and simulate real-world scenarios. The implications stretch far beyond gaming or virtual reality. It covers logistics, healthcare and finance, as organizations map physical locations, simulate autonomous interactions and collect rich multimodal data reflecting the messiness of lived experience. xAI, for instance, has reportedly recruited AI researchers from Nvidia—one of the pioneers in world model development through its Omniverse platform that simulates entire environments for robotics and industry. According to Nvidia's own estimates, the market potential for world models could approach the size of the global economy itself,

illustrating the enormous expectations around these systems. Musk has even indicated plans for xAI to launch an AI-generated game powered by such models, signaling the pace of advancement in real-world simulation and interaction.

This change signifies a shift: advancement is evaluated by both linguistic competence and the capacity to reason, anticipate outcome and align actions with observable events. Should this new generation of models succeed, it may significantly reduce the disparity between computational processes and real-world adaptation, representing a substantial advancement comparable to earlier transformative technologies and further strengthening the connection between theoretical knowledge and practical application.
Google DeepMind's Genie 3 generates video frame by frame, Meta's V-JEPA learns passively by observing raw video and Nvidia's Omniverse simulates entire physical environments for robots and manufacturing. These models are not just content generators, but blueprints for next-generation agency. AI systems that could navigate highways, interact in hospitals and build virtual spaces indistinguishable from reality.

Industrial revolution 2.0 — AI in payments and beyond

It's tempting to draw parallels between AI's emergence and previous flashes of life changing breakthrough such as the steam engine, electrification, digital computing. But AI's novelty is its ubiquity and pace: it's everywhere at once, transforming industries collectively rather than sequentially. The impact on finance is especially profound, as the sector both shapes and reflects broader changes in trust, risk and exchange.

AI is already reconstructing finance, replacing static rule-based systems with dynamic, learning algorithms. In payments, this 'second Renaissance' is most vividly felt. AI-driven fraud detection works in milliseconds, learning from billions of transactions and thwarting attacks faster than any team of analysts could manage. Payment orchestration becomes an active process, not merely a switchboard but an intelligent conductor who is routing transactions based on real-time data, historical behavior and evolving risk landscapes.

Cross-border transactions, once hobbled by cost and delay, now move with the fluidity of local payments, as smart AI algorithms navigate the thicket of exchange rates, compliance rules and settlement protocols. Consumer-facing experiences are being tailored in ways unimaginable a decade ago, with AI parsing individual spending patterns and dynamically adjusting risk assessment and product recommendations.

The numbers underpin the sensation: global digital payment transaction value is projected to reach nearly $15 trillion by 2027, with AI fueling this growth not just by efficiency, but by enabling entirely new business models e.g. predictive commerce, micro-transactions and autonomous finance.

Embodied intelligence and the fate of finance

What makes this shift "unprecedented" is the emergence of economic intelligence as a sort of ambient infrastructure. No longer passive executors of human intent, payment systems powered by AI increasingly anticipate, adapt and even out-think their human creators. Previous revolutions—coins, credit cards, databases—served as extensions of human will. AI is different: it learns, evolves and sometimes makes decisions before we even realize there's a choice to be made.

The **Financial Times** report further notes that these embodied AI models, designed to perceive physics and cause-effect relationships, represent a major leap from today's video-generation models such as OpenAI's Sora, which only predict image frames. World models instead understand how objects interact in space and time, allowing AI to make sense of the physical context it operates within.

The Financial Times observes leading institutions investing in embodied world models, pushing AI out of the virtual space and into the physical. The implications for finance are profound: payments might soon be managed not just by algorithms parsing digital data, but by agents interacting in the real world making sense of spatial, sensory and behavioral cues on the fly. The boundary between digital money and human experience grows thinner, hinting at a future where commerce is woven directly into the environments we inhabit.

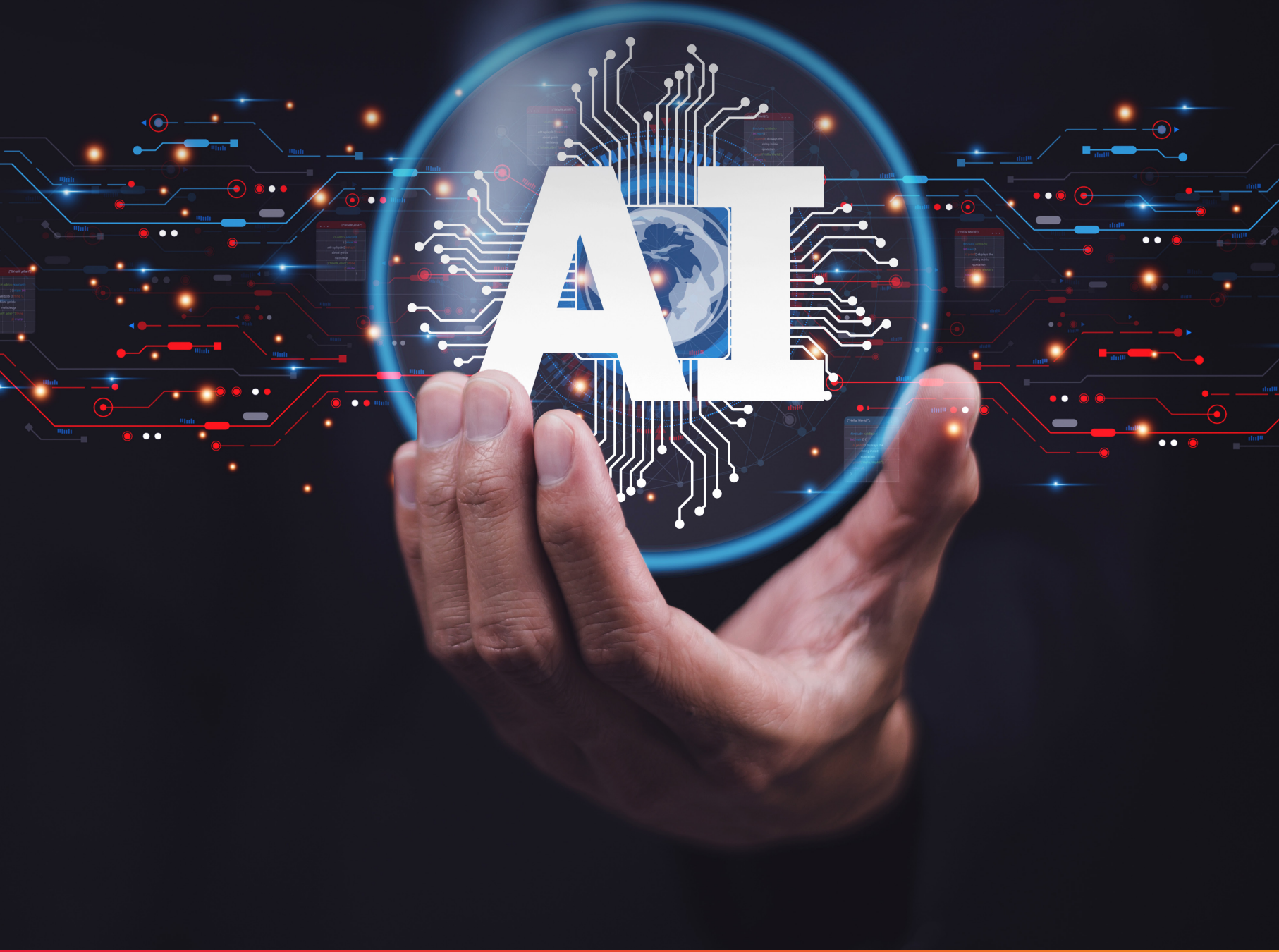
Toward a new map of commerce and creativity

Perhaps the most provocative prospect is what this means for human creativity and commerce writ large. If AI world models deliver on their promise, the next generation of payment platforms will not simply move money, they will enable entirely new ways to imagine, construct and interact with value itself. Payment and financial infrastructure become less an administrative "pipe" and more an intelligent participant—able to see trends, foresee risks and even negotiate on behalf of users.

This, finally, is why AI's arrival in finance should be considered a Renaissance rather than a mere evolution. Like Gutenberg's press and the internet,

artificial intelligence has begun to redraw the systems of participation, agency and imagination. The smartest payments platforms won't just transact; they will sense, learn and make creative choices at scale. The result: economic and creative activity that is more adaptive, more inclusive, and, perhaps, a step closer to the fluid intelligence we have long reserved for humans alone.

In this latest chapter of technological history, the question is no longer what AI can do for finance, but how finance and commerce more broadly will be remade by an intelligence capable of operating both within and beyond the boundaries of human experience. With the age of world models upon us, transformation is set not only for industry but for the very architecture of possibility itself. As AI adoption in financial services accelerates, the landscape will shift beyond transactions to immersive experiences. The emergence of the experiential economy means consumers increasingly demand seamless, personalised and emotionally resonant journeys, where payments cease to be mere invisible utilities and instead become central to the experience. AI acts as the bridge, elevating payments from a back-end process to an intelligent, adaptive interface that anticipates needs, enhances convenience and even delights with tailored offers or contextual advice. In this evolving world, every payment is no longer just a transfer of value; it becomes a curated touchpoint, an experience shaped by real-time insights and deep learning, making finance an active participant in crafting memorable, meaningful moments for individuals and businesses alike. The future of payments, powered by AI, will thus be defined not only by speed and security, but by the ability to transform each payment into a unique and valuable experience.



02 AI Use Case – Smart Budgeting
by Sandhya

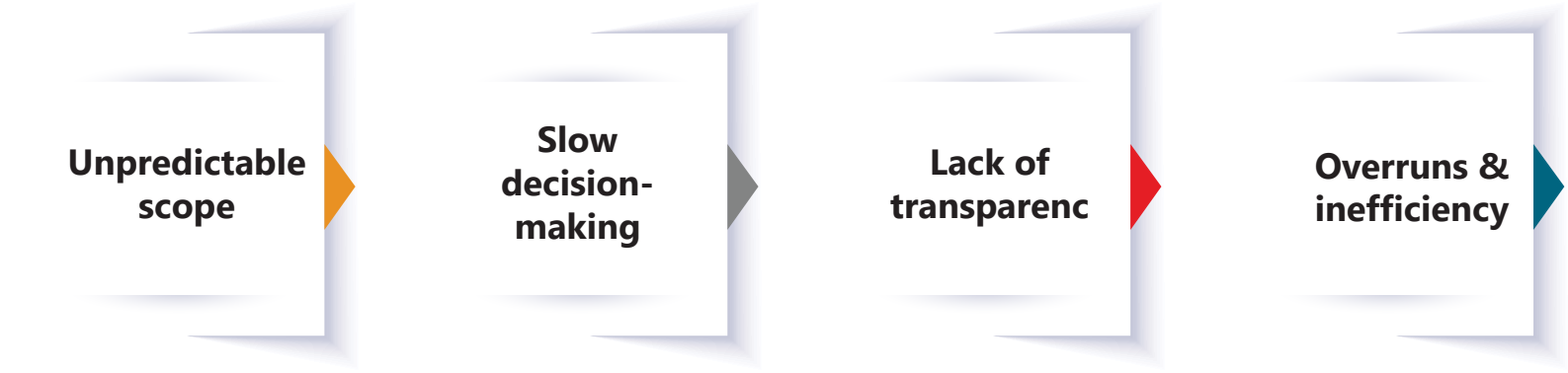
Sandhya holds 13 years of experience in designing and delivering cloud-native applications across diverse domains including Banking & Financial Services, Computer Telephony Integration (CTI), Stock Trading, and Tax Calculation. She has extensively worked on Monolithic, Microservice and SOA architecture. She also has experience in building CI/CD pipelines using Git, Jenkins and implementing SRE principles for high available and scalable systems.

Software projects are difficult to budget. Shifting requirements, unpredictable timelines, and rapidly changing technologies often cause projects to overrun budgets or under-deliver. Traditional budgeting models are rigid, annual/quarterly, and top-down.

Hence, struggle to keep pace with such dynamic environments.

AI is transforming the way businesses manage budgets by enabling **data-driven micro-budgeting.** This approach breaks down large budgets into smaller, more agile allocations and uses predictive analytics to continuously adjust spend. The result is better control, faster decisions, and improved project outcomes.

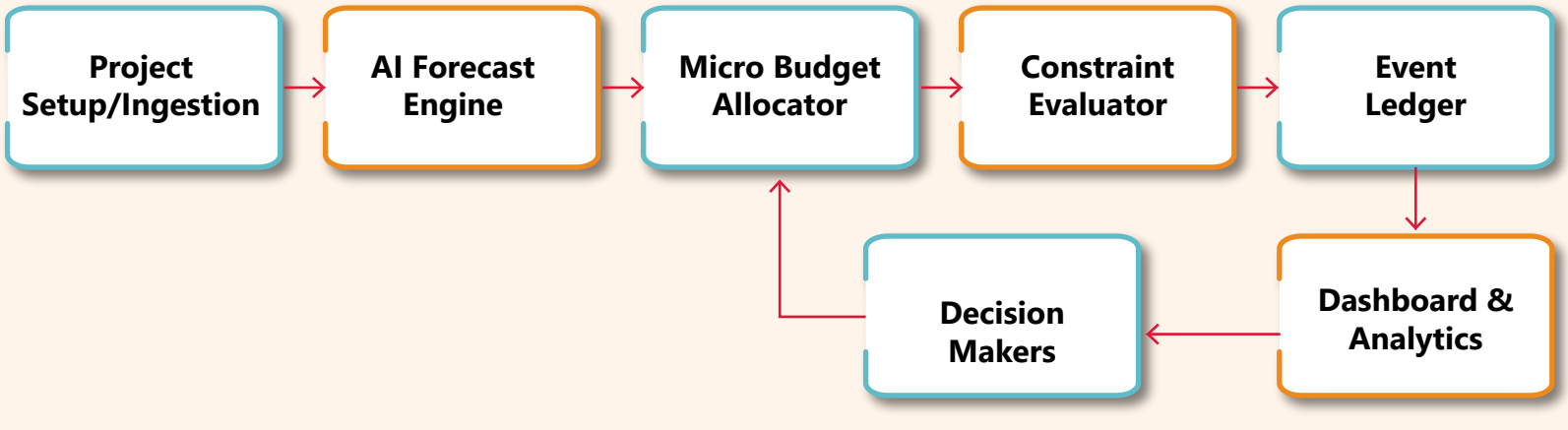
Problems with Traditional Budgeting in Software Projects



AI-Powered Micro-Budgeting Works by enhancing the budgeting process:

1. Forecasting Spend Patterns: Machine learning models predict resource consumption based on historical project data.
2. Dynamic Allocation: AI reallocates funds to critical modules or teams in real time.
3. Variance Detection: Automated alerts trigger when actual spend deviates from planned spend.
4. Risk Mitigation: Predictive models flag high-risk areas (e.g., modules prone to scope creep) early.

This transforms budgeting from a static plan into a living system that learns and adapts.



Component Responsibilities

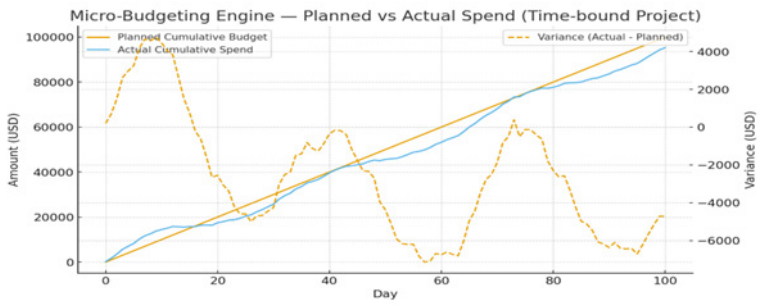
- **Ingestion Service:** Validates incoming project definitions, micro-budgets, time bounds, and constraints (hard/soft). Emits well-formed events to the Budget Engine.
- **Budget Engine:** Orchestrates allocation, re-allocation, and enforcement. Keeps business logic versioned. Emits ledger entries for every allocation and spend.
- **Micro-Budget Allocator:** Applies allocation algorithms (e.g., prorated by phase, priority queues, rule-based reserves) to split macro budgets into micro-budgets tied to time windows.
- **Constraint Evaluator:** Evaluates policies (e.g., min headroom, max daily burn, vendor caps) and produces suggested adjustments or enforcement actions.
- **Budget Manual (Event Store):** append-only store of all budget events (allocated, adjusted, reserved, spent). Enables deterministic replays.
- **Analytics & Reporting:** Materialized views for dashboards, forecast models, and KPI calculations.
- **Realtime Rules Engine:** Enforces live checks and triggers alerts when thresholds are crossed.

Performance Improvement in Action

The performance chart illustrates the financial progression of a 100-day time-bound project by comparing the planned cumulative budget against the actual cumulative spend, using synthetic (sample) data. The planned budget line represents how much was intended to be spent over time, while the actual spend line shows the real expenses incurred day by day. The variance, calculated as the difference between the two, highlights whether the project is staying within budget, overspending, or underspending at any point. This type of visualization is crucial for monitoring project health, enabling early interventions, and demonstrating how closely forecasts or AI-generated budgets align with actual financial performance.

Performance Graph

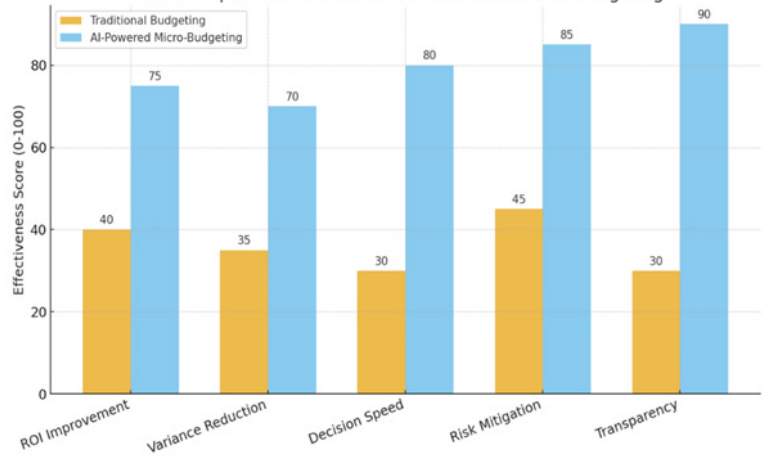
This chart demonstrates how micro-budgeting smooths variance and reduces overspending compared to traditional budgeting.



Business Impact

- **Improved ROI:** Continuous reallocation ensures funds go where they create the most value.
- **Agility:** Rapid adjustments align with agile software delivery cycles.
- **Risk Reduction:** Early detection of variance reduces financial exposure.
- **Enhanced Trust:** Transparent reporting builds confidence among stakeholders.

KPI Comparison: Traditional vs. AI-Powered Micro-Budgeting



Future Perspective

As AI models mature, they will: - Incorporate natural language processing to interpret project updates automatically. - Using reinforcement learning to optimize budget decisions in complex, multi-team projects. - Integrate with DevOps pipelines to tie spending directly to delivery outcomes.

AI-driven micro-budgeting is reshaping how software projects are funded and managed. By combining predictive analytics, real-time monitoring, and adaptive allocation, businesses can overcome traditional budgeting challenges. The result is faster delivery, controlled costs, and higher ROI—a win for both project teams and leadership.

03 AI in Engineering: Finding the brilliant middle
by Swagata

Swagata is a technology leader with 18+ years of experience in mobile app architecture, digital banking solutions, and enterprise product development. Currently he is heading product development at Comviva and drives the delivery of mobiquity® Pay apps (Android, iOS, Web).

He thrives at the intersection of technology, product innovation, and team empowerment, building platforms that scale globally and deliver exceptional customer experiences.

AI has moved my work up-stack. The busywork shrank, and the thinking got sharper. Here's how it actually plays out for me.

Where AI Makes Life Easier (and Faster)

Code generation as scaffolding, not truth. When I'm staring at a blank file, AI gives me a skeleton to start with. It's the framing of a house, not the interior design. I still adjust or finetune for perfection. We must treat AI output like a junior developer's draft; promising, but reviewable.

Architecture review on tap. AI is a surprisingly good "rubber duck" with guardrails. I feed it an ADR, and it spots brittle couplings, hidden fan-out, or something so hidden in layers that it is difficult to find at a glance. It's especially handy for "what could go wrong?" prompts. AI helps highlight brittle couplings; but only if the ADR and architecture context are rich. Garbage in, garbage out.

Test cases without the canvas of pain. For unit and property-based tests, AI drafts broad coverage quickly: happy paths, boundary values, locale mishaps, and concurrency edges.

Code review that never gets tired. Linting is one thing, but AI adds context. It flags not just "style" but the likely runtime cost of a piece of code, the risks in exception swallowing, and the missing cancellation in coroutines. I don't accept everything, but it pushes me to justify choices. It writes crisp review comments I can edit, not rewrite.

Bug triage with a memory for patterns. Given a failure log, AI proposes likely culprits. It scans our code for the issue and can provide solutions to discuss and choose from. I often get three to four hypotheses; one is usually close enough to shorten the hunt.

Documentation that people actually read. AI converts rough notes into a readable "How It Works" with diagrams, ADR summaries, and onboarding checklists. It helps maintain parity between code and docs by generating diffs of behaviour ("what changed?") rather than just prose. This alone saves enough time to pay for the tool.

The Over-Dependence Trap

AI can fabricate APIs, invent config flags, or cite a non-existent RFC - confidently. If we don't verify against primary docs, we'll ship fiction. No, it's not malicious; it's probabilistic. So it is always important to verify whatever is returned from AI.

"Quick" snippets might pull in non-compliant code or suggest insecure defaults. Humans must be the last mile on threat models and compliance. Question the solution, finetune it till we get what we want.

If we accept "typical" patterns blindly, we get cookie-cutter systems that don't fit our latency, data shape, or regulatory constraints. Wisdom is knowing when to break the pattern.

Why a Human in the Loop Is a Must

Trade-offs are human. Only a human can weigh correctness vs. time-to-market, or vendor lock-in vs. team expertise. AI can list trade-offs; a human must choose them.

Ethics and accountability. When a decision impacts privacy, fairness, or safety, "the model said so" won't fly. Someone has to own the call and sign their name under it. Though it is generated by AI, a human is the person who curated it. So the responsibility lies on the human.

Learning is not optional. A human-in-the-loop mindset forces deliberate practice: read the diff, run the benchmark, inspect the trace. We improve while AI accelerates, not instead of us. AI becomes a team member, not our replacements.

My Working Rules (hard-earned, not theoretical)

Write the spec first. Even a short “goal, constraints, non-goals” note keeps AI on-task and makes reviews meaningful.

Use AI for breadth, keep humans for depth. AI explores options; I finalise one with numbers (latency, memory, cost).

Treat AI like a sharp tool. Anything security-sensitive gets manual threat modelling and dependency checks.

Automate the verification. Pair AI with tests, linters, SAST/DAST, and canaries. If it’s not tested, it doesn’t exist.

Keep a knowledge spine. Consolidate final decisions in ADRs and runbooks; don’t let knowledge live only in chats.

My Take

AI is like amplification. If our process is sloppy, AI will help us create more, faster, slop. If our process is disciplined, AI compresses cycle time and boosts quality. My experience: the teams that win are the ones that **own the prompt, the process,** and the proof; they ask precise questions, run tight feedback loops, and verify with data.

It is important to define what is rewarded. For example, if we reward test case coverage without adding rails for quality, AI will happily write surface level tests just to increase coverage.

AI made my life simpler by clearing the path: less boilerplate, faster reviews, broader test nets, better docs. But the destination is still mine to choose and to justify. Keep the human in the loop, and keep shipping things you can defend in daylight.

04 Fintech's Competitive Edge: AI as the Differentiator
by Gopal

Gopal is a results-driven Engineering Manager with deep expertise in microservices architecture, distributed systems, and event-driven platforms building large-scale FinTech and mobile wallet solutions. With a passion for technology leadership, he is an explorer at heart, diving into the latest AI advancements including LLMs, RAG, and Agentic AI etc. He combines this curiosity with hands-on experience in microservices, distributed systems, and mobile wallet platforms to drive impactful innovation.

AI is not an add-on in fintech — it is the new data layer and decisioning fabric. Over the last decade fintech systems moved from batch/statistical models to real-time, LLM-enabled, retrieval-augmented and agentic pipelines that power fraud detection, real-time scoring, conversational banking, compliance automation and programmatic trading. This article explains the core technical primitives (LLMs, embeddings, vector DBs, RAG, agentic orchestration), how they are applied in fintech, production patterns (AI/MLOps, latency SLAs, monitoring), and how to build robust, auditable systems. Key references and vendor/tech callouts are included for engineers who want to implement these patterns.

Core primitives fintech teams must master

Large Language Models (LLMs)

LLMs (transformer-based models) provide fluent natural-language understanding and generation: prompt conditioning, few-shot learning, fine-tuning, and adapter-based domain specialization (FinLLMs). In fintech, LLMs power natural questioning (e.g., "why was my wire blocked?"), contract summarization, regulatory text extraction, and code generation for internal automation. Institutional deployments often require custom fine-tuning or supervised re-training on proprietary transaction and policy data to ensure domain fidelity and regulatory safety.

Embeddings & Vector Databases

Embeddings convert text, transactions, or documents into dense vectors used for semantic search and similarity. Vector databases (Pinecone, Milvus, Weaviate, etc.) store and retrieve these vectors at scale and millisecond latencies, which is the backbone of RAG systems and context retrieval for conversational agents. Vector stores also enable nearest-neighbour analytics (e.g., clustering suspicious transaction patterns) and are commonly used alongside metadata filters (time, account id, geo).

Retrieval-Augmented Generation (RAG)

RAG couples semantic retrieval from authoritative internal sources with an LLM for grounded generation — e.g., pull the last 60 transactions + policy text, then have the LLM explain reasons for a decline. In finance this reduces hallucinations and gives auditable evidence: the LLM's output is backed by retrieved text chunks. RAG is now a canonical architecture in fintech for knowledge-centric tasks.

Agentic AI & Orchestration

Agentic systems are composed of autonomous agents that can plan, call tools, and act across services (KYC agent, Fraud investigation agent, Payments settlement agent). Frameworks like LangChain, AutoGen and multi-agent patterns are already used to orchestrate multi-step flows, e.g., gather docs → run OCR & biometric match → query watchlists → decide payout. Agentic models are not "AGI"; they are orchestrated LLMs with tool use and memory.

Explainability & XAI

Finance demands explainability. SHAP (SHapley Additive exPlanations) is a widely used framework to explain feature contributions for individual predictions (credit score, fraud risk) and is now standard in regulated model deployments. Combine SHAP with counterfactual explanation tooling and you get both feature-level and scenario-level interpretability required by auditors/regulators.

Real-World Use Cases that Matter

Fraud detection & risk scoring — RAG + streaming embeddings

Problem: detect evolving fraud patterns in real time across millions of payment events.

Pattern: streaming ingestion → feature store → embeddings + vector similarity → anomaly scoring model (ML/DL) → agentic triage.

Pipeline specifics:

- Ingest transactions into Kafka/Kinesis; compute streaming features & embeddings; persist to feature store and upsert vectors into a vector DB (Pinecone/Milvus).
- Use a hybrid detector: (a) a supervised model for known patterns + (b) an unsupervised or contrastive model that detects novel anomalies. The supervised model uses tabular & sequence features (RNN/transformer for session sequences); the anomaly detector compares recent transaction embedding to historical clusters.
- For high-risk alerts, spawn an investigation agent that pulls supporting evidence (device fingerprint, geolocation, recent chargebacks) via RAG and primes an LLM to summarize the likely attack vector for human investigators.
  Production constraints: latency often required <50ms risk scoring on payment rails for soft declines / allow/deny decisions (Mastercard's Decision Intelligence reports sub-50ms scoring in production systems). Use fast inference stacks and canary traffic routing.

Conversational & contextual banking — LLM + RAG + vector DBs

Problem: customers need precise, contextual answers that reference account history, product terms, and regulatory constraints.

Pattern: user query → canonicalization → embedding retrieval (accounts docs, T&Cs, transaction history) → context assembly → LLM answer with citations (RAG).

Key engineering points: maintain per-customer context vectors (privacy constraints), implement prompt templates that require the LLM to output source pointers (document id, snippet offsets), and enforce guardrails to block policy-sensitive outputs. Use domain-tuned LLMs and local model instances where regulatory constraints require data residency.

Credit underwriting & explainability (XAI + ensemble models)

Problem: automate instant lending decisions while meeting fairness and regulatory requirements.

Pattern: structured + alternative data ingestion (transactions, bills, telco), feature augmentation, ensemble scoring (tree ensembles + neural models), SHAP for per-loan explanation.

Key details: produce an "explainability payload" per decision (SHAP values with top contributing features, confidence bands, and recommended human review triggers). Use thresholder human-in-loop for borderline cases and maintain audit trails.

RegTech: automated compliance and policy change monitoring

Problem: keep track of regulatory changes and ensure product compliance.

Pattern: automatic ingestion of regulatory text (web scrapers + document loaders) into vector stores → periodic RAG reports that extract obligations and map them to internal controls → ticket generation in GRC systems. This reduces compliance triage time by large factors and surfaces obligations proactively.

Robo-advisory & portfolio construction — LLM + time-series models + RAG

Problem: deliver hyper-personalized wealth plans and explanations at scale.

Pattern: client profile + goals → RAG fetch of policy constraints and market data → optimization (convex/QP or RL policy) → LLM to generate explanation and next actions. Use a hybrid stack: numerical optimizer (fast) + LLM for human-facing narratives and compliance language.

AML / Transaction surveillance — graph analytics + RAG for case writing

Problem: detect and document money-laundering rings across networks.

Pattern: ingest transaction graphs → apply graph ML (GCNs, node2vec) for suspicious cluster detection → augment with RAG to auto-draft SAR (Suspicious Activity Report) narratives using retrieved evidence. Graph ML exposes entities and clusters; the drafted narrative comes from RAG that references supporting transactions. This significantly reduces case analyst time while preserving auditability.

Synthetic data & privacy-aware modeling

Problem: lack of usable labeled data due to privacy and compliance constraints.

Solution: generate high-fidelity synthetic transaction datasets for model development, stress testing and scenario generation. Synthetic frameworks preserve statistical properties while protecting PII; Nvidia's synthetic data consolidations show how enterprise players are investing in this capability. Use synthetic data for edge test cases and algorithmic robustness checks, but always validate on real holdout sets for distributional fidelity.

Production patterns: design, deployment & governance

AI/ML Ops & model lifecycle

- Model packaging: containerize models (Seldon Core, BentoML) for consistent deployment.
- CI/CD for models: automated unit & integration tests, shadow & canary deployments, A/B experiment orchestration. Seldon Core and similar frameworks are commonly used to manage production ML traffic and A/B experiments.
- Observability: telemetry for latency, throughput, accuracy, drift, hallucination rate (RAG mismatch metrics) and cost metrics. Implement automated retraining triggers based on data/label drift.

Latency & scaling

- Use model quantization, distillation and on-host caching to ensure sub-100ms inference for scoring. For high throughput (payment rails), lightweight model ensembling or constraint-based models reduce decision time to the required SLA (e.g., <50ms for many card processing flows).

Governance, audit, and compliance

- Produce per-decision artifacts: model version, input snapshot, feature vector, SHAP explanation, retrieved evidence (for RAG), and LLM prompt + response. Store these artifacts in an immutable audit store for regulatory review. Regular bias audits, explainability reviews, and stress tests are mandatory in production fintech environments.

The Emerging Frontier

Agentic AI orchestration (multi-step autonomous flows)

What it is: collections of specialized agents (retriever, policy agent, verifier, executor) that collaborate to complete end-to-end tasks.

Fintech example: autonomous loan origination agent — it (a) ingests an application, (b) runs identity verification (OCR + face match), (c) executes credit scoring, (d) negotiates offer terms, (e) executes payout. Agents communicate through structured messages (JSON tool APIs) and a coordinator enforces transaction safety and audit logging. Frameworks like LangChain, AutoGen and similar libraries provide primitives for tool-calling, memory, and orchestration.

Engineering considerations:

- **Tool abstraction interface** (OpenAPI/JSON schema) to call downstream services (KYC, payments, ledger).
- **Planner vs. Executor split:** Planner composes steps, Executor calls tools with strict rate limits and RBAC.
- **Safety layer / circuit breaker:** human review threshold for financial value > X or when model confidence < Y.
- **Observability:** traceability of each agent call, inputs, embeddings, and LLM responses for compliance.

Generalist & multimodal models for finance

New generalist models handle tabular data, text and images in a single architecture — enabling workflows that combine document OCR, numerical analysis and narrative generation in one pass (e.g., read a bank statement image → extract balances → update portfolio). This reduces pipeline complexity and improves latency for multistep tasks.

On-device & privacy-first fintech (TinyML + federated learning)

For biometric verification and offline risk checks, TinyML models and on-device inference preserve latency and data residency. Federated learning enables model improvements across devices without moving raw PII — helpful for mobile banking in privacy-sensitive jurisdictions.

Quantum-accelerated finance (research stage)

Quantum algorithms (QAOA, quantum annealing) and quantum-inspired optimizers are being investigated for portfolio optimization and risk simulation. It's an R&D frontier that fintechs with quant desks and research budgets are exploring. Expect hybrid classical →quantum pipelines for specific optimization tasks.

Risks, mitigations, and operational guardrails

Hallucinations: Always combine LLMs with retrieval and grounding. For high-value decisions require deterministic checks (rule-based validators) after generation.

Model drift/data leakage: Automate drift detection, refresh embeddings, and reindex vector stores frequently. Prevent leakage by strict data governance controls and tokenization of PII.

Bias / fairness: Use SHAP and fairness metrics (demographic parity, disparate impact) and deploy fairness-aware reweighing if required.

Operational risk & systemic failure: limit autonomous agent actions with hard limits, require two-party confirmations for high-value flows, and maintain human-in-the-loop for exceptions.

Implementation playbook: step-by-step (for a fintech team)

Phase 0 — Discovery & safety baseline

- Inventory data sources, map regulatory requirements, take a security/privacy baseline.
- Establish KPIs: fraud false positive rate, time-to-decision, model latency, eNPS for customer support.

Phase 1 — Minimal RAG pilot

- Choose a narrow use case (e.g., "account closure disputes"). Build a RAG pipeline: ingest docs, embed, vector DB, LLM with coached prompts, SLAs for human review. Validate outputs and measure hallucinations and accuracy.

Phase 2 — Harden & productionize

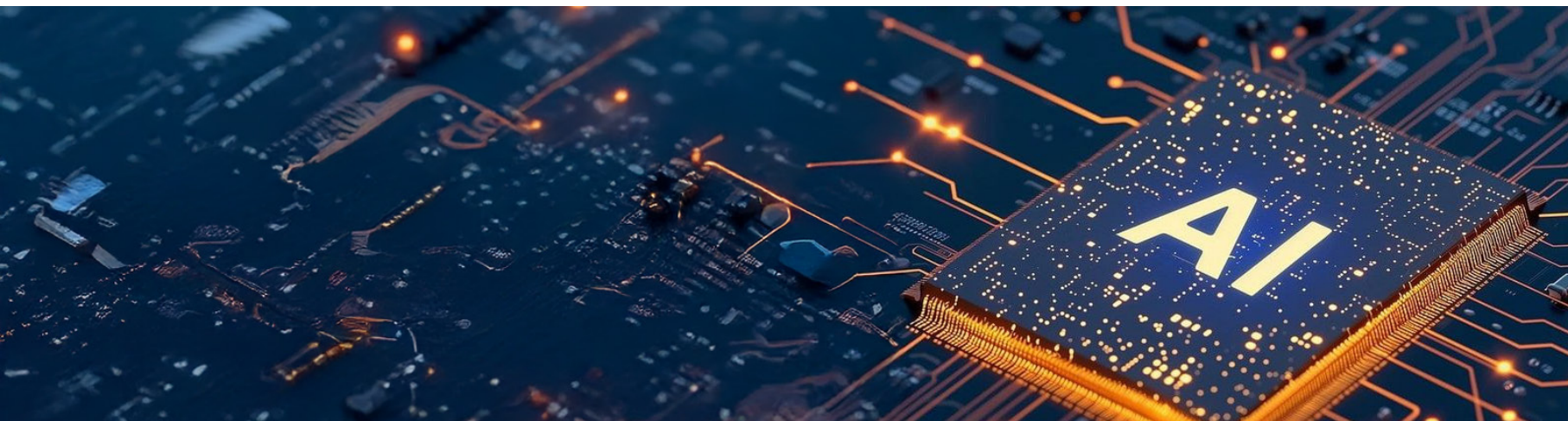
- Containerize models, integrate Seldon/MLflow/Kubeflow for lifecycle, add monitoring, define retrain cadence, and implement canary rollouts.

Phase 3 — Scale & govern

- Expand to high-value flows, deploy agentic orchestration for multi-step processes, implement model governance & audit pipelines, run fairness and stress tests, and maintain a model catalog with owners and SLAs.

Closing: design principles for excellent fintech AI

- **Ground everything:** LLMs must be grounded with retrieval and deterministic validators for monetary decisions.
- **Build observability by design:** artifacts for every decision are required for trust and compliance.
- **Human + AI governance:** maintain human oversight thresholds and clear escalation paths.
- **Iterate safely:** start narrow, measure, harden, then scale.



Building the Future of Fintech, Responsibly

The promise of AI in fintech isn't just about efficiency; it's about building a more intelligent, responsive, and secure financial ecosystem. The technical primitives and architectural patterns discussed in this article — from LLMs and vector databases to agentic orchestration and advanced MLOps — are the building blocks.

However, technology alone isn’t enough. True innovation in this space requires a commitment to **grounded AI, transparent governance, and a human-in-the-loop approach.** As these systems become more autonomous, their auditable and explainable nature will be the ultimate differentiator, ensuring trust and compliance in an industry where safety is paramount.

By prioritizing these principles, we can move beyond the hype and create robust, ethical, and transformative AI systems that redefine the future of finance.

05 The AI revolution is here – What is your take?
by Ved

Ved has around 13 years of experience in the IT industry, with over a decade in the fintech domain. He has led multiple successful deployments of fintech products across regions in Asia, the Middle East, and Africa. He is passionate about technology and keenly follow how Artificial Intelligence is transforming our daily lives and the way we work, always exploring ways to leverage these advancements for real-world impact.Ved has around 13 years of experience in the IT industry, with over a decade in the fintech domain. He has led multiple successful deployments of fintech products across regions in Asia, the Middle East, and Africa. He is passionate about technology and keenly follow how Artificial Intelligence is transforming our daily lives and the way we work, always exploring ways to leverage these advancements for real-world impact.

Introduction

Indeed, AI is no longer just a buzzword—it's here, it's real, and it's transforming the world around us. As software professionals and product organizations, if we are not looking seriously at AI today, then that's our problem. Ignoring it is like trying to ignore an elephant standing right next to us.

For years, "AI" and now "AI agents" have been used as hot buzzwords in conferences and boardroom discussions. But today, they are much more than hype. According to PwC, AI could contribute $15.7 trillion to the global economy by 2030, with financial services and retail being among the biggest beneficiaries. Already, 65% of financial institutions have implemented AI-driven solutions for fraud detection, customer support, and process automation.

AI's Impact on Fintech

The impact is visible: AI is improving productivity, accuracy, debugging, issue resolution, and real-time insights across industries. Imagine a system that not only alerts you to potential issues before they occur but also understands user behavior and adapts to their needs. That is the empathy AI brings—systems that not only function but feel like they truly "take care" of the people using them.

Coming from the fintech domain with nearly a decade of experience on digital payments, I strongly believe embedding AI in payment ecosystems is not just an option—it's a necessity. In 2024 alone, digital wallets accounted for 64% of global e-commerce payments and 38% of point-of-sale payments (Worldpay Global Payments Report).

By using AI to understand problems in advance, we can ensure seamless payment services, which directly improves customer satisfaction and drives revenue growth. But AI in fintech is not limited to reliability—it's also about personalization.

Confidential Studies show that 80% of consumers are more likely to engage with companies that personalize experiences. By analyzing customer behavior for each service used, AI can enable new pricing strategies, such as adaptive service rates or subscription-based models. Instead of a "one-size-fits-all" approach, financial products can become tailored experiences, making users feel valued and understood.

Imagine a fintech app that isn't just a tool but a partner, helping you plan your budget, monitor spending habits, and suggest smarter financial decisions. That's not just technology—it's empathy, delivered at scale. This shift is already starting, and it's only going to accelerate.

Global Payments Trends – Worldpay Global Payments Report

Figure 1: E-commerce Payment Method Shares (2024 vs 2030)

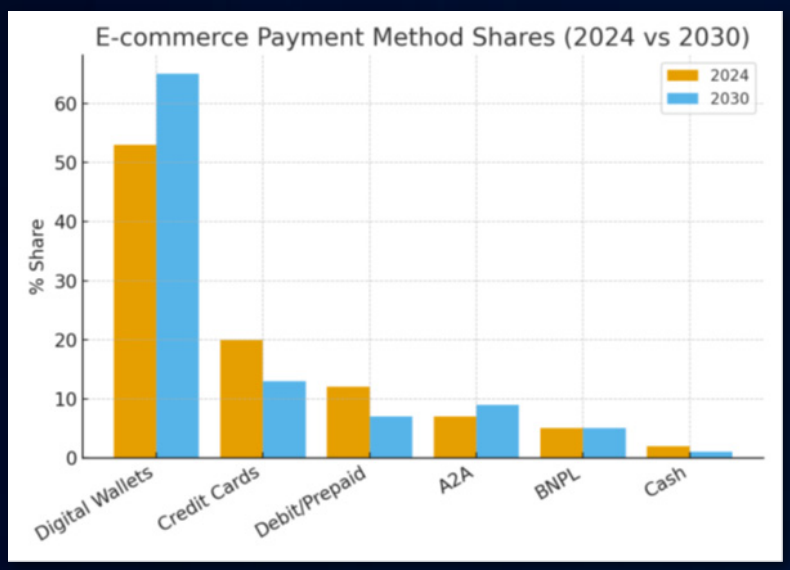


Figure 2: POS Payment Method Shares (2024 vs 2030)

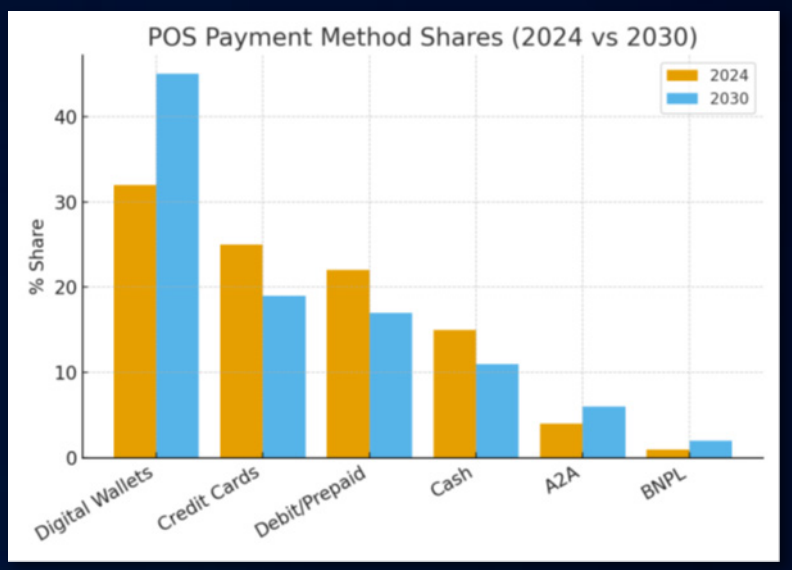
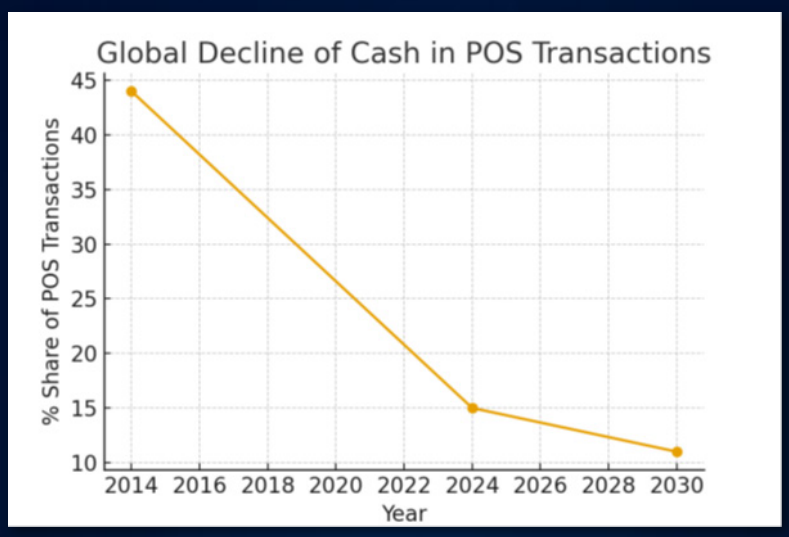


Figure 3: Global Decline of Cash in POS Transactions (2014–2030)



Conclusion

The AI revolution is not about replacing humans with machines. It's about building systems that can augment human decision-making, bring empathy into digital experiences, and make products feel more human than ever before. For fintech, it's about moving from being a payment processor to becoming a financial partner.

The AI revolution is here—and the data is clear. Those who embrace it will shape the future of finance; those who ignore it will simply be left behind.

06 Grafana with AI
by Ajit

Ajit has a vast experience that spreads around cybersecurity and building secure and scalable solutions for REST API's as per OWASP standards.
He is currently working on java, spring boot and microservices development and is excited to build new creative FinTech solutions.

Turning Monitoring into Business Intelligence

What is Grafana?

Grafana is an open-source observability and monitoring platform that allows businesses and IT teams to visualize, analyze, and understand data in real time. It's widely used to track metrics, logs, and traces from infrastructure, applications, and business systems. Originally popular for time-series dashboards, Grafana has now evolved into a full-stack observability solution with integrations spanning databases, cloud platforms, monitoring tools, and AI-driven forecasting. Grafana is built to be a pluggable observability platform. Its power lies in connecting to multiple data sources, exporters, and backends.

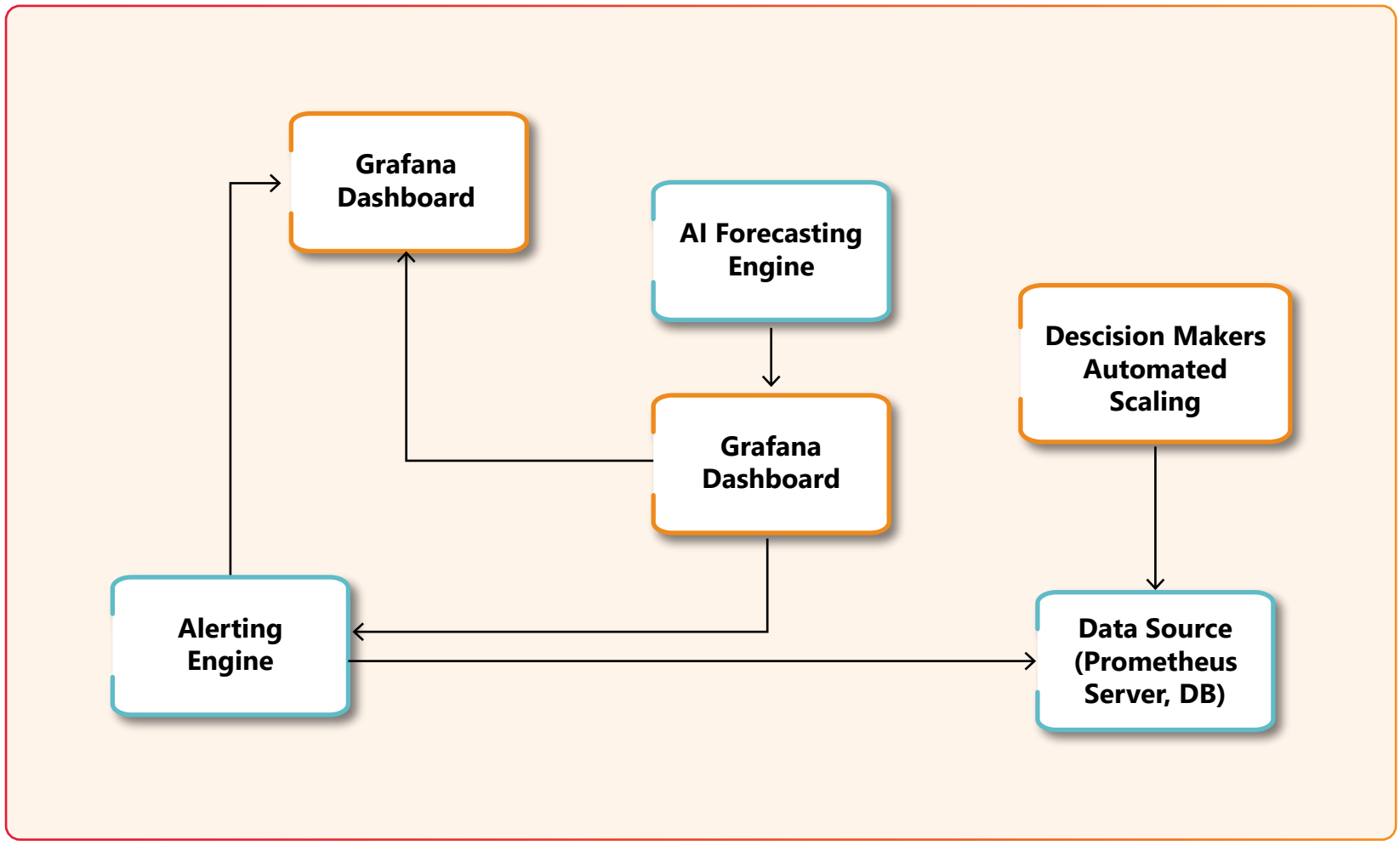
Key Features of Grafana

- **Customizable Dashboards &**
- **Multi-Source Integration Support**
- **Authentication &**
- **Real-Time Monitoring**
- **Alerting System**
- **K6, Load Testing Tool**
- **Scalable & Extensible**

How Grafana works

Grafana works with an AI forecasting engine by combining real-time observability with predictive intelligence. Grafana collects metrics, logs, and traces from multiple sources and visualizes them through dashboards, while the AI forecasting engine processes this time-series data using machine learning models to detect patterns, trends, and anomalies. The engine then generates predicted metrics and risk scores, which are fed back into Grafana panels alongside real-time data. This integration enables teams not only to monitor current system health but also to anticipate failures, optimize resource usage, and prevent costly downtime.

Infrastructure KPIs	**Application Performance KPIs**	**Business KPIs**	**Reliability & Forecasting KPIs (AI + Grafana)**
CPU Utilization	Response Time / Latency	Revenue per Transaction / Order	Risk Scores
Memory Usage	Error Rates	Customer Engagement Metrics	Cloud Spend Forecasts
Disk I/O & Storage Usage	Throughput (Requests per second)	Conversion Rates	Anomaly Detection Alerts
Network Throughput & Latency	Service Dependencies Health	Operational Costs	Cloud Spend Forecasts
Uptime / Availability	Transaction Success Rate	SLA Compliance	Predicted System Failures



How Business Uses Grafana

IT Operations (ITOps)
Monitor infrastructure uptime & performance.

DevOps & SRE Teams
Observe application health, latency, and errors.

Cloud Cost Monitoring
Visualize and forecast cloud spend.

Business Intelligence
Track KPIs like revenue, sales, or customer activity.

MSPs (Managed Service Providers)
Provide clients with SLA and performance dashboards

Grafana KPIs (Key Performance Indicators)

Root Cause Analysis (RCA):

Logs + Metrics + Traces can be correlated using AI.
AI highlights dependencies across microservices.
Cuts down time wasted during incident triage.

Example: Instead of 100 alerts, Grafana AI groups them into 1 “root cause cluster.”

Natural Language Queries (NLQ):

AI enables chat-like querying inside Grafana:
Instead of writing PromQL/SQL, ask: “Show me the average latency for service X in the last 24 hours.”
Great for non-technical users like business analysts and project managers.

Predictive Alerts:

Traditional alerts = “CPU > 80%”.
AI-driven alerts = “CPU expected to reach 90% in the next 15 minutes based on trend.”
This shifts monitoring from reactive to predictive.

07 AI Agents in Java: Automating Non-Functional Requirements for Secure, Scalable Apps *by Mohit*

Mohit is a seasoned professional with over 16 years of experience around mobile banking, mobile money, financial services, digital recharge solutions and digital payments. He loves to work towards solutions which provide seamless end-user experiences and deliver better value to clients. Mohit enjoys reading about new technologies and trends in the industry.

Non-functional requirements (NFRs) such as test coverage, security, and performance are critical in enterprise Java applications. Unlike functional features, these are harder to validate manually. AI copilots like Cursor and Claude are increasingly being used in DevOps pipelines to enforce NFRs automatically, improving quality and security without adding overhead.

Code Coverage

Java teams often use **JaCoCo** or **Cobertura** for coverage tracking. Cursor and Claude enhance this process by:

- **Analyzing reports** to identify untested or high-risk modules.
- **Generating JUnit 5 and Mockito tests** automatically. Cursor can draft parameterized unit tests, while Claude suggests test scenarios derived from specifications or user stories.
- **Prioritizing risky code paths** by correlating recent Git changes with historical bug density.

Static Application Security Testing (SAST)

SAST tools like SonarQube, Checkmarx, or Fortify catch vulnerabilities in source code. AI copilots make these scans more effective by:

- **Automating scans in CI/CD pipelines** to catch issues before merge.
- **Reducing noise** by classifying vulnerabilities based on severity and exploitability. Cursor flags unsafe patterns in code, while Claude explains the risks in plain English for developers.
- **Proposing fixes** such as replacing unsafe SQL concatenation with parameterized **PreparedStatements.**

Dynamic Application Security Testing (DAST)

DAST tools like **OWASP ZAP** and Burp Suite analyze applications at runtime. AI copilots strengthen these tests by:

- **Generating intelligent attack payloads** to probe Spring Boot REST APIs and web endpoints.
- **Interpreting runtime anomalies** — Claude summarizes scan results, while Cursor assists in applying patches.
- **Focusing scans** on sensitive endpoints, such as authentication or payment workflows.

Performance & Scalability

Performance is often overlooked during early development but remains a vital NFR. Cursor and Claude help by:

- **Driving load tests** using **JMeter** or **Gatling,** with realistic traffic patterns.
- **Analyzing JVM metrics** such as garbage collection cycles, heap usage, and thread activity. Claude highlights bottlenecks, while Cursor recommends code or configuration optimizations.
- **Suggesting tuning parameters** for JVM, Spring Boot thread pools, or database connections to

AI in NFRs at a Glance

NFR Area	Traditional Tools	AI Copilot Role
Code Coverage	JaCoCo, Cobertura	Cursor writes missing tests, Claude suggests scenarios
SAST	SonarQube, Checkmarx	Cursor flags unsafe code, Claude explains vulnerabilities
DAST	OWASP ZAP, Burp Suite	Cursor applies patches, Claude interprets scan results
Performance	JMeter, Gatling	Cursor optimizes configs, Claude analyzes logs

Unified Benefits

By combining coverage tools, security scanners, and performance testers with copilots like Cursor and Claude, Java teams gain:

- **Faster feedback loops** across multiple NFRs.
- **Reduced manual effort** through automated test generation and fix suggestions.
- **Better compliance** with audit-ready reports that satisfy regulatory needs.

Conclusion: AI copilots such as **Cursor and Claude** are transforming Java development by actively supporting non-functional requirements. They generate tests, guide security scans, interpret results, and optimize performance — helping teams build applications that are not just functional, but also secure, efficient, and scalable.

Recent Awards

★ **mobiquity® Pay - Best Digital Wallet (Platinum)** ★
Future Digital Award by Juniper Research

★ **mobiquity® Pay – Financial Inclusion** ★
FSTech Award 2025

★ **mobiquity® One – OL - Best Payment Facilitator** ★
Drivers of Digital Award 2025

In the Media

Embedded Payments
Addressing critical challenges in $25 billion space
with Srini