

# The Synthetic & Hybrid Data Universe

A Master Catalog — from Data Modalities to the Systems that Generate Them

**TIER 1 · DATA LAYER** Enterprise · Scientific · Knowledge · Imaging · Audio

**TIER 2 · SYSTEMS LAYER** Digital Worlds · Decision Engines · AI Models · Twins · Email Agents

**SYNTHETIC · REAL · HYBRID** Pure-synthetic by default; real and hybrid (real + synthetic) where a real anchor adds value.

Pradeep · XpertSystems.ai · Bellevue, WA  
pradeep@xpertsystems.ai · 206-245-6747

## How to Read This Catalog

This catalog is organized into two tiers. The **Data Layer** enumerates what we generate — datasets and artifacts across every modality. The **Systems Layer** enumerates what generates and consumes that data — the simulators, decision engines, and models that emit artifacts and act on them. A single engagement often spans both: a digital world (Systems Layer) emits hundreds of enterprise artifacts (Data Layer), which then train a decision engine (Systems Layer) that produces new labeled outcomes (Data Layer).

|                                                                                                                                           |                                                                                                                                    |
|-------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|
| <b>Data Layer</b><br>WHAT we generate. Datasets and artifacts across every modality — the deliverables a model trains, tests, or runs on. | <b>Systems Layer</b><br>WHAT generates and consumes it. The simulators, decision engines, and models that emit data and act on it. |
|-------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|

Each entry pairs a type with concrete, production-ready examples so a use case maps directly to a deliverable.

## Synthetic, Real & Hybrid — Three Ways to Source a Dataset

Every deliverable in this catalog can be sourced three ways. **Pure synthetic** is machine-generated and privacy-safe by construction — our default, carrying no real PII or PHI. **Real** data is licensed from partner organizations and de-identified before any use. **Hybrid** blends a de-identified real anchor with synthetic scale — the real records preserve authentic correlations, rare events, and messiness that generators smooth over, while synthetic provides volume, class balance, and edge-case coverage. Typical hybrid blends run from a light 10% real / 90% synthetic anchor up to 40% / 60% for high-fidelity use cases; the exact ratio is set empirically per task. Where a modality below benefits from a real anchor, a **Hybrid option** line notes what real data adds.

**TIER 1**

**Data Layer**

*What we generate — datasets and artifacts across eight modalities, available pure-synthetic, real, or hybrid.*

## 1. Enterprise & Business Documents

Realistic business paperwork with correct fields, layouts, cross-references, and internally consistent entities.

| Data Type   | Concrete Examples / Artifacts                                                                              |
|-------------|------------------------------------------------------------------------------------------------------------|
| Procurement | Purchase orders (PO), requisitions, RFQs/RFPs, vendor quotes, three-way match packets, goods-receipt notes |

| Data Type            | Concrete Examples / Artifacts                                                                        |
|----------------------|------------------------------------------------------------------------------------------------------|
| AP / AR              | Invoices, credit notes, statements, remittance advice, dunning letters, aging reports                |
| Finance & Accounting | General ledgers, trial balances, P&L, balance sheets, cash-flow statements, expense reports, budgets |
| Contracts & Legal    | MSAs, NDAs, SOWs, employment & lease agreements, amendments, redlines, clause libraries              |
| HR & People          | Offer letters, resumes/CVs, performance reviews, org charts, payroll registers, benefits enrollment  |
| Sales & Marketing    | Proposals, order forms, pricing sheets, CRM records, lead lists, campaign briefs, case studies       |
| Compliance & Audit   | SOX narratives, audit workpapers, KYC/AML files, policy docs, risk registers, SOPs                   |
| Operations           | Shipping manifests, bills of lading, packing slips, work orders, maintenance & inspection logs       |
| Executive & Planning | Board decks, roadmaps (PPTX), OKR trackers, business cases, meeting minutes, status reports          |

## 2. Structured & Transactional Data

Row-level tabular and event data with referential integrity, realistic distributions, seasonality, and controllable anomalies. A prime candidate for hybrid: a de-identified real anchor sharpens fraud signals and true distributions.

| Data Type               | Concrete Examples / Artifacts                                                                                                                                                                                      |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Financial Transactions  | Card payments, ACH/wire, ledger postings, trades/orders, fraud-labeled event streams<br><b>Hybrid option:</b> <i>a de-identified real transaction anchor captures genuine spend patterns and fraud signatures.</i> |
| Insurance Claims        | P&C, workers' comp, health/group claims, adjudication histories, subrogation records<br><b>Hybrid option:</b> <i>real claim histories add authentic severity and adjudication variance.</i>                        |
| Retail & E-Commerce     | Orders, carts, SKUs, inventory movements, returns, pricing history, clickstream, loyalty                                                                                                                           |
| Telecom & IoT           | CDRs, network telemetry, sensor time series, smart-meter reads, device event logs                                                                                                                                  |
| Supply Chain            | Demand forecasts, shipments, warehouse movements, supplier lead times, BOMs, routing                                                                                                                               |
| Manufacturing           | Machine sensor streams, defect/QC records, OEE metrics, IIoT protocol traffic<br><b>Hybrid option:</b> <i>real fault and defect records supply rare failure signatures synthetic under-represents.</i>             |
| Energy & Grid           | Load/demand curves, generation mix, wholesale prices, outage events, renewable output                                                                                                                              |
| Web & Product Analytics | Sessions, funnels, A/B events, cohort/retention tables, feature-usage logs                                                                                                                                         |
| Graph & Relational      | Customer 360, entity resolution, fraud rings, comm graphs, knowledge graphs                                                                                                                                        |

### 3. Scientific & Research Data

Physically and statistically plausible measurements for R&D, simulation, and model pre-training. Real experimental data can anchor hybrid sets where measurement noise and structure are hard to generate.

| Data Type             | Concrete Examples / Artifacts                                                                                                                                                                              |
|-----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Genomics & Omics      | DNA/RNA (FASTQ/VCF), variant calls, expression matrices, proteomics, microbiome<br><b>Hybrid option:</b> <i>de-identified real omics data anchors models where biological structure resists synthesis.</i> |
| Clinical Trials       | Synthetic cohorts, CRFs, adverse-event logs, dosing/PK-PD curves, lab panels<br><b>Hybrid option:</b> <i>de-identified real trial trajectories add authentic longitudinal signal.</i>                      |
| Chemistry & Materials | Molecular structures (SMILES), reactions, spectra (NMR/IR/MS), material properties                                                                                                                         |
| Physics & Engineering | Simulation outputs (CFD/FEA), particle events, signal traces, control telemetry                                                                                                                            |
| Earth & Climate       | Weather series, satellite rasters, hydrology, seismic, air-quality, emissions                                                                                                                              |
| Astronomy             | Light curves, spectra, synthetic sky surveys, transient/event catalogs                                                                                                                                     |
| Biology & Ecology     | Population dynamics, species observations, epidemiological spread models                                                                                                                                   |
| Social Science        | Survey responses, census-like microdata, economic indicators, mobility panels                                                                                                                              |

### 4. Medical & Healthcare Data

Privacy-safe patient data spanning records, imaging, and device streams — pure-synthetic and HIPAA-safe by default, or hybrid where a de-identified real anchor is licensed and de-identified before use.

| Data Type              | Concrete Examples / Artifacts                                                                                                                                                                       |
|------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EHR / Clinical Records | Encounters, problem lists, med orders, notes, discharge summaries, ICD/CPT-coded records<br><b>Hybrid option:</b> <i>de-identified real EHR anchors real comorbidity and care-pathway patterns.</i> |
| Medical Imaging        | X-ray, CT, MRI, ultrasound, PET, mammography, DICOM studies with radiology reports<br><b>Hybrid option:</b> <i>de-identified real studies add rare-finding coverage generators miss.</i>            |
| Pathology & Lab        | Whole-slide images, histopathology, blood panels, cultures, genetic tests                                                                                                                           |
| Cardiology & Signals   | ECG/EKG, EEG, echocardiograms, Holter streams, vitals telemetry<br><b>Hybrid option:</b> <i>real signal recordings capture authentic waveform variability.</i>                                      |
| Claims & RCM           | Medical/pharmacy claims, prior-auth, EOBs, denial/appeal records, utilization mgmt                                                                                                                  |
| Specialty Datasets     | Oncology, neurology, cardiology, endocrinology, respiratory, rare-disease cohorts                                                                                                                   |
| Wearables & Remote     | Continuous glucose, heart rate, sleep, activity, remote-monitoring streams                                                                                                                          |

| Data Type     | Concrete Examples / Artifacts                                                |
|---------------|------------------------------------------------------------------------------|
| Public Health | Disease surveillance, immunization registries, outbreak/epidemic simulations |

## 5. Images & Visual Documents

Rendered and generated imagery with ground-truth labels, masks, and OCR-ready text layers. Real labeled imagery can anchor hybrid sets where defect or scene diversity is hard to synthesize.

| Data Type             | Concrete Examples / Artifacts                                                                                                                                                  |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Business Documents    | Scanned POs, invoices, receipts, checks, forms, IDs, tax forms — with OCR annotations                                                                                          |
| Medical Imaging       | X-rays, CT/MRI slices, ultrasound, retinal scans, dermatology photos, DICOM output                                                                                             |
| Industrial & Defect   | Manufacturing defects, PCB/weld inspection, surface anomalies, thermal imagery<br><b>Hybrid option:</b> <i>real defect images add the long tail of real-world anomalies.</i>   |
| Autonomous & Robotics | Driving scenes, LiDAR/depth maps, grasp/manipulation frames, perception sets<br><b>Hybrid option:</b> <i>real sensor logs anchor rare driving and manipulation edge cases.</i> |
| Geospatial            | Satellite/aerial imagery, land-use segmentation, building footprints, change detection                                                                                         |
| Retail & Product      | Shelf/planogram images, product photos, barcode/label crops, try-on imagery                                                                                                    |
| Security & Biometric  | Synthetic faces, fingerprints, document-fraud samples, privacy-safe surveillance                                                                                               |
| Charts & Diagrams     | Synthetic charts, tables, floor plans, engineering drawings, handwriting                                                                                                       |

## 6. Audio, Speech & Conversations

Training-safe synthetic audio and dialogue with speaker turns, timestamps, transcripts, and intent/emotion labels. Multi-engine TTS (Kokoro + Piper) for licensing-clean voice. Consented, de-identified real audio can anchor hybrid sets for acoustic and conversational realism.

| Data Type            | Concrete Examples / Artifacts                                                                                                                                                                  |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Healthcare Dialogue  | Doctor–patient consults, triage calls, nurse hand-offs, telehealth, med counseling<br><b>Hybrid option:</b> <i>de-identified real consult audio adds authentic clinical conversation flow.</i> |
| Customer Service     | Support/contact-center calls, IVR flows, complaint handling, retention/collections<br><b>Hybrid option:</b> <i>de-identified real call logs anchor genuine customer language and intent.</i>   |
| Meetings & Business  | Stand-ups, board/exec meetings, sales calls, interviews, negotiations, all-hands                                                                                                               |
| Financial & Advisory | Advisor–client calls, loan interviews, claims intake, compliance disclosures                                                                                                                   |
| Voice Assistants     | Command/intent utterances, wake-words, multi-turn task dialogues, barge-in                                                                                                                     |

| Data Type                | Concrete Examples / Artifacts                                                                                                                                                         |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Training & Audiobooks    | Compliance narration, e-learning, audiobook chapters, safety briefings                                                                                                                |
| Speech Recognition (ASR) | Accented/noisy/multilingual speech, far-field audio, code-switching, jargon<br><b>Hybrid option:</b> <i>consented real recordings add true accent, noise, and acoustic diversity.</i> |
| Ambient & Non-Speech     | Machine sounds, environmental audio, alarms, acoustic anomaly sets                                                                                                                    |

## 7. Knowledge Bases & Text Corpora

Structured knowledge and natural-language corpora for RAG, search, and LM training — internally consistent and citation-linked.

| Data Type            | Concrete Examples / Artifacts                                                        |
|----------------------|--------------------------------------------------------------------------------------|
| Enterprise Knowledge | Wikis, SOPs, runbooks, FAQ banks, help-center articles, policy libraries, onboarding |
| Q&A & Instruction    | Question–answer pairs, instruction–response sets, multi-hop QA, grounded RAG         |
| Domain Corpora       | Legal, medical, financial, technical text with terminology and citations             |
| Structured Knowledge | Ontologies, taxonomies, knowledge graphs, entity/relation triples, glossaries        |
| Communications       | Emails, Slack/Teams threads, tickets (Jira/ServiceNow), chat logs, forums            |
| Code & Technical     | Source files, docstrings, commit histories, bug reports, API docs, configs           |
| Multilingual         | Parallel corpora, localized content, cross-lingual QA, transliteration sets          |
| Exam & Certification | Licensing banks (SIE/Series), assessment items with rationales and validators        |

## 8. Workflows & Process Data

Process-level data capturing how work actually flows through an enterprise — the observable trace of operations.

| Data Type             | Concrete Examples / Artifacts                                                            |
|-----------------------|------------------------------------------------------------------------------------------|
| Job-to-Task Graphs    | Role → task → action decompositions across legal, HR, sales, IT, mfg, retail, healthcare |
| Enterprise Workflows  | SAP/Oracle/Salesforce/ServiceNow process flows, approvals, exceptions, hand-offs         |
| Process Mining        | Event logs (XES), variant/conformance data, bottleneck and rework patterns               |
| Long-Horizon Projects | Multi-phase project spines, sprint-indexed events, milestone & dependency graphs         |

### TIER 2

#### Systems Layer

*What generates and consumes the data — worlds, engines, and models. These systems can be trained and evaluated on synthetic, real, or hybrid data.*

## 9. Digital Worlds & Enterprise Twins

Deterministic discrete-event simulators that model an entire enterprise and emit hundreds of cross-linked, internally consistent artifacts. These are the engines that produce Data-Layer output at scale.

| System Type               | What It Is / Produces                                                                              |
|---------------------------|----------------------------------------------------------------------------------------------------|
| Enterprise World Twins    | Reconstructive enterprise simulators emitting Slack, Jira, email, docs, decks, spreadsheets        |
| Operating-Model Twins     | Typed graph twins with drivers, scenarios, and configuration dials yielding many model variants    |
| Project & Clinical Twins  | World-agnostic resolver layers with sprint-indexed event spines for projects and care pathways     |
| Multi-Connector Platforms | Hub-and-spoke world backbones spanning many industries and enterprise connectors                   |
| Sector Operating Twins    | Vertical world models for healthcare, financial services, supply chain, government, telecom, legal |

## 10. Decision Engines & Intelligence

Systems that consume data — synthetic, real, or hybrid — and emit decisions, each decision paired with its outcome, producing a new supervised dataset. Both a consumer and a producer of data; training on a hybrid input mix improves fidelity on real-world decisions.

| Engine Type              | Decisions & Outputs                                                                  |
|--------------------------|--------------------------------------------------------------------------------------|
| Underwriting & Risk      | Insurance underwriting, credit risk, exposure scoring with decision → outcome labels |
| Claims Adjudication      | Automated claim decisions, denial/approval logic, subrogation and fraud flags        |
| Utilization Management   | Prior-authorization, medical-necessity, care-pathway decisions with rationale traces |
| Prioritization & Routing | Triage, queue prioritization, assignment, and escalation decision datasets           |
| Decision Intelligence    | Approval workflows, policy-driven decisions, what-if and counterfactual outcomes     |

## 11. Decision Learning Packets & Trajectories

Two parallel product lines that package how decisions are made into training-ready data. A Decision Learning Packet (DLP) captures a single decision point in full — inputs, policy, rationale, and outcome; a Decision Learning Trajectory (DLT) captures the connected chain of decisions across a long horizon. Both share one production spine: deterministic seeding, spec-driven validators, Grade A+ gates, an IEEE XES process layer, and branded brochures. Packets and trajectories can be grounded in synthetic scenarios or in de-identified real decision records for hybrid fidelity.

| Product Line                         | What It Delivers                                                                                                                            |
|--------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Decision Learning Packets (DLP)      | Richly-labeled single-decision packets — inputs, policy, rationale traces, and outcome — built on a proven 5–8 step packet architecture     |
| Decision Learning Trajectories (DLT) | Long-horizon decision sequences (state → decision → outcome → next state) with rationale traces and success/failure labels, parallel to DLP |

| Product Line             | What It Delivers                                                                                                                                       |
|--------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| Domain Coverage          | Insurance underwriting, workers' comp, property/auto claims, credit disputes, prior auth, transplant allocation, triage, and SAR filing                |
| Finance & Agentic Series | Trajectory data spanning equity trading, options, portfolio allocation, risk, and long-horizon agentic planning, tool-use, and multi-agent negotiation |
| Process Layer            | IEEE XES event logs with variant and conformance structure, so each packet or trajectory carries an auditable process trace                            |
| Validation & Gates       | Spec-driven validators (13-check suites) and Grade A+ quality gates ( $\geq 92/100$ ) with SHA-256 substream seeding for full reproducibility          |

## 12. AI Models & Evaluation Systems

The trained systems shipped as deliverables, plus the harnesses that measure them — where the data becomes a model, and the model is put under test. Models can be fit on synthetic, real, or hybrid corpora depending on the fidelity the task demands.

| System Type               | What It Delivers                                                                                                                                                                                                                                     |
|---------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Trained Models            | Domain models fit on synthetic, real, or hybrid corpora; fine-tunes; task-specific classifiers and predictors<br><b>Hybrid option:</b> <i>a hybrid corpus blends de-identified real records with synthetic scale for higher real-world fidelity.</i> |
| Agent Evaluation          | Multi-turn agent eval platforms across SAP, Oracle, Salesforce, ServiceNow and others                                                                                                                                                                |
| Eval Frameworks           | Contamination-resistant evaluation kits, rubrics, adversarial and benchmark suites                                                                                                                                                                   |
| Archetype Libraries       | Reusable archetype sets (e.g., 10 per domain) as repeatable expansion units for coverage                                                                                                                                                             |
| Self-Verifying Validators | Question banks and datasets with built-in validators enforcing Grade A+ quality gates                                                                                                                                                                |

## 13. Coding-Task Factory & RL Environments

Deterministic factories that generate long-horizon, executable coding tasks as reinforcement-learning environments — each task a runnable repo with an automated grader, dense reward, and captured solve trajectories. Verifiable by construction: baselines provably fail, reference solutions provably reach full reward, and every task is reproducible from a seed and gated to Grade A+ before it ships.

| System Type               | What It Delivers                                                                                                                                  |
|---------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| Coding-Task Factory       | Deterministic, NumPy-seeded factory that generates verifiable coding tasks and packages only those passing the Grade A+ gate                      |
| RL Environments           | Sandboxed task instances exposing read/write/list/run actions with a hidden-test grader, dense partial-credit reward, and binary success          |
| Long-Horizon Coding Tasks | Bug-fix-from-issue, feature-from-spec, and repo-wide refactor tasks requiring many dependent steps to a verifiable terminal state                 |
| Finance-Domain Tasks      | Quant tasks — returns compounding, drawdown, position sizing, order accounting — whose correctness requires domain knowledge to author and verify |
| Solve Trajectories        | Full action/observation/reward/done sequences per task, exported as structured JSON for RL training and SFT on successful solves                  |

| System Type       | What It Delivers                                                                                                                            |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Automated Graders | Hidden test suites as ground truth; baselines provably fail and reference solutions provably reach full reward, with reward-spread evidence |
| Quality Gate      | Five-dimension A+ scorer ( $\geq 92/100$ ): baseline discrimination, reward spread, determinism, structure, and grader execution            |

## 14. Reasoning & Agent Systems

Generators of cognition-level data — the step-by-step reasoning and tool-use trajectories used to train and align autonomous systems.

| System Type                 | What It Generates                                                                               |
|-----------------------------|-------------------------------------------------------------------------------------------------|
| Engineered Reasoning Traces | Chain-of-thought corpora for cybersecurity, finance, law, IP, risk, medicine — verifiable steps |
| Agent Trajectories          | Multi-turn tool-use logs, API-call sequences, task-to-action traces with success/failure labels |
| Voice-Agent Systems         | Long-horizon conversational agents and voice-platform interaction data                          |
| Reasoning Corpora           | Design-debate, failure-scenario, and decision-rationale traces from long-horizon engagements    |

## 15. Enterprise World Compiler — Graph, MCP & Agent Evaluation

The layer that turns an emitted digital world into an AI-ready, queryable, callable asset. The Compiler ingests a connected enterprise (Item 9) and produces a canonical entity catalog, a cross-system knowledge graph, a live MCP server agents can call, and a verified tool-calling evaluation dataset — each stage gated and reproducible. Where Item 9 emits artifacts, the Compiler makes them traversable, tool-callable, and benchmark-ready.

| Component                     | What It Delivers                                                                                                                                                                                                                                                                                                                                               |
|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Enterprise Knowledge Graph    | Deterministic compiler that resolves five spokes (warehouse, chat, M365, Jira, MCP/SaaS) into one canonical entity catalog and cross-system graph, with an 8-check validation gate (unique IDs, two-way edge resolution, zero orphans, catalog↔graph consistency).                                                                                             |
| Live MCP Server               | The world served as callable tools over Model Context Protocol — search every entity type, retrieve full detail, and traverse relationships across systems in a single call. Installs as a desktop extension or endpoint; every result grounded in the validated graph, no fabricated IDs.                                                                     |
| Verified Tool-Calling Dataset | Benchmark-grade agent-eval data: each item is a multi-hop question, the correct tool-call sequence, and a ground-truth answer re-derived from the graph and independently verified. Seven multi-hop archetypes (root-cause, blast-radius, PHI-exposure, cross-system traverse, ...), emitted in OpenAI, Anthropic, and MCP formats.                            |
| Hard Negatives                | Adversarial items proven to be genuine traps — plausible-but-wrong answers and insufficient-data cases — that test whether an agent rubber-stamps plausible output or hallucinates when data is missing. The property that turns a training set into a benchmark.                                                                                              |
| EnterpriseBench (held-out)    | A formal benchmark built on the verified traces: 15,000 items across three verticals, with a sealed held-out test set (questions shipped, answers withheld so buyers cannot overfit), deterministic scoring (set-F1 for ID lists, exact-match for scalars and verdicts), and a cross-vertical generalization split — train on two industries, score on a third |

| Component                     | What It Delivers                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                               | the agent has never seen. Every answer key is graph-verified; the benchmark is content-hashed and versioned.                                                                                                                                                                                                                                                                                                                                                                                                                                                   |
| Enterprise Search & Retrieval | A hybrid full-text index (BM25) over every document, ticket, chat message, and entity name in the world — 1.3M+ searchable items on the full Healthcare world. The differentiator: every search hit carries its graph IDs, so a result is an entry point into the knowledge graph, not a dead end. Search → graph → traverse. Deterministic, dependency-free, with a validation gate that confirms every hit resolves to a real node.                                                                                                                          |
| GraphRAG Package              | Retrieval-augmented generation over the graph. Where flat RAG returns isolated chunks, GraphRAG returns connected subgraphs — incident-centered communities (e.g. 4,000 communities on the full Healthcare world) that give an LLM structured, connected context. Ships community summaries, graph-aware retrieval, and 5,000 community-level evaluation questions with graph-verified answers (5,000/5,000 verified, zero mismatches).                                                                                                                        |
| Vertical-Agnostic Engine      | The same seeded, connected, validated pipeline — graph, MCP server, and eval dataset — compiles a complete Enterprise World for any of 14 verticals, with industry-specific entities and system-of-record semantics swapped in and identical linkage and validation guarantees. Demonstrated on three verticals to date; the discovery-based engine adds new verticals without code changes.                                                                                                                                                                   |
| Proof Point — 3 Verticals     | Reproduced at production scale across three structurally-different verticals from one engine, with zero code changes: Healthcare & Life Sciences (357,547 nodes / 1,928,360 edges), Automotive (357,547 / 2,104,416), and Business & Professional Services — consulting, legal, accounting (357,546 / 1,919,151). Each passed all 8 validation gates with zero orphans and yielded 5,000 verified tool-calling traces (4,250 positives with 0 answer mismatches, 750 hard negatives) — every positive answer independently re-derived from that world's graph. |

## 16. AI Email Agents & Deployable Agent Chassis

A reusable email-native agent framework — receive, parse, process, reply, log — where the processing step is a swappable module. Each module is a deployable SKU, delivering instant, structured replies inside the channel users already work in.

| System Type            | What It Delivers                                                                                                                                                                                           |
|------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Agent Chassis          | Reusable receive → parse → process → reply → log framework — fast polling, de-duplication, self-reply loop guard, timezone-aware handling, and logged replies inherited by every module                    |
| Deterministic Modules  | Fixed input → computation → output SKUs carrying proprietary logic; e.g. signal classification over up to 20 inputs with a structured reply in roughly 15 seconds                                          |
| LLM-Powered Modules    | Claude-driven modules with a built-in safety layer — best-judgment replies, scope and guardrail enforcement, and polite decline on out-of-scope or unsafe requests; a new SKU is often a new system prompt |
| Enterprise Use Cases   | Support-inbox triage & routing, invoice and document extraction, order and status lookups, compliance and document checks, and internal knowledge Q&A                                                      |
| Delivery & Integration | Email-native today, API-native on request; deployable on managed infrastructure with logging, so agents fit existing workflows without a new app to adopt                                                  |

## How the Two Tiers Connect

The layers form a closed loop. A Systems-Layer world simulates an enterprise and emits Data-Layer artifacts; those artifacts — synthetic, or hybrid with a de-identified real anchor — train a Systems-Layer decision engine or model; the engine's decisions and the model's evaluations become new Data-Layer records; and those feed the next iteration. This is why XpertSystems can deliver either a dataset, a simulator, a decision engine, a trained model, or an evaluation harness — they are stages of one pipeline.

- **Worlds & twins** EMIT artifacts (Tier 2 → Tier 1).
- **Models & engines** TRAIN ON artifacts — synthetic, real, or hybrid (Tier 1 → Tier 2).
- **Engines & evals** PRODUCE new labeled data (Tier 2 → Tier 1).

## Cross-Cutting Generation Standards

- Deterministic engines with SHA-256 substream seeding for full reproducibility.
- Grade A+ quality gates ( $\geq 92/100$ ) with 6-seed stability sweeps.
- Controllable statistical fidelity, edge cases, drift, and injected anomalies.
- Ground-truth labels, schemas, and validators shipped with every deliverable.
- Synthetic output is licensing-clean and privacy-safe by construction (no real PII/PHI). Hybrid and real datasets use partner-licensed data that is de-identified before any use, with re-identification review and scoped, contracted permitted-use terms.
- Delivered across CSV, JSON, DICOM, WAV, DOCX, XLSX, PPTX, and PDF.

**XpertSystems.ai** — *From the data your models need to the systems that generate it. Synthetic by default, hybrid where a real anchor adds fidelity.*