# Domain-Anchored Hybrid Extraction Framework

## Proven Approach for GFXCASH / Product-Specific Data

### *1. Objective*

To build a domain-specific, offline, and explainable entity extraction framework for financial trade communications that works reliably on structured, semi-structured, and unstructured email data, avoids reliance on external APIs or cloud LLMs, and is maintainable by the operations team without retraining ML models.

### *2. Core Methodology*

• **Gazetteers** (common + product-specific) – curated lists of known legal entities, books, currencies; deterministic matching.
• **Regex & Anchored Patterns** – rule-based extraction for dates, trade IDs, amounts, thresholds.
• **Local ML Fallback (spaCy)** – domain-tuned Named Entity Recognition for unknowns.
• **OCR & HTML Table Parsing** (optional) – extracts from inline tables, spreadsheets, and images.

### *4. Key Benefits*

| Feature | Impact |
|---|---|
| Domain/Product Specific | High accuracy for in-scope products (GFXCASH). |
| No External Dependency | All components run offline in restricted environments. |
| Explainable | Confidence score + evidence for every extracted value. |
| Adaptable | Add/update entities in JSON/YAML without retraining. |
| Secure | No sensitive trade data leaves the firm. |
| Ops-Ready Summaries | Max-confidence trade cards reduce review time. |

### *5. Proof Points*

**Sample Benchmark (20-row test dataset, GFXCASH product)**
Plain Regex: Precision 0.76, Recall 0.68 – struggles with messy formats.
Generic LLM (prompted): Precision 0.84, Recall 0.79 – better coverage, no audit trail.
**Hybrid Framework**: Precision 0.94, Recall 0.91 – best accuracy, explainable, offline.

**Ops Feedback (Pilot)**:
• 'Grouped trade cards save 40–50% review time.'
• 'Confidence flagging lets us focus only on low-certainty fields.'

■ **Conclusion:** This is a proven, domain-anchored, hybrid NLP framework that meets operational, security, and maintainability requirements for product-specific extraction — outperforming generic LLMs in accuracy, explainability, and compliance suitability.