Receipt OCR API

Extraction structurée de données depuis des reçus scannés

FastAPI · Groq Cloud · EasyOCR · PaddleOCR · Unstructured · Docker

Février 2026

Contexte & Objectifs

Problème

Extraire automatiquement des informations structurées depuis des **reçus scannés** (PDF) en JSON normalisé.

Objectifs

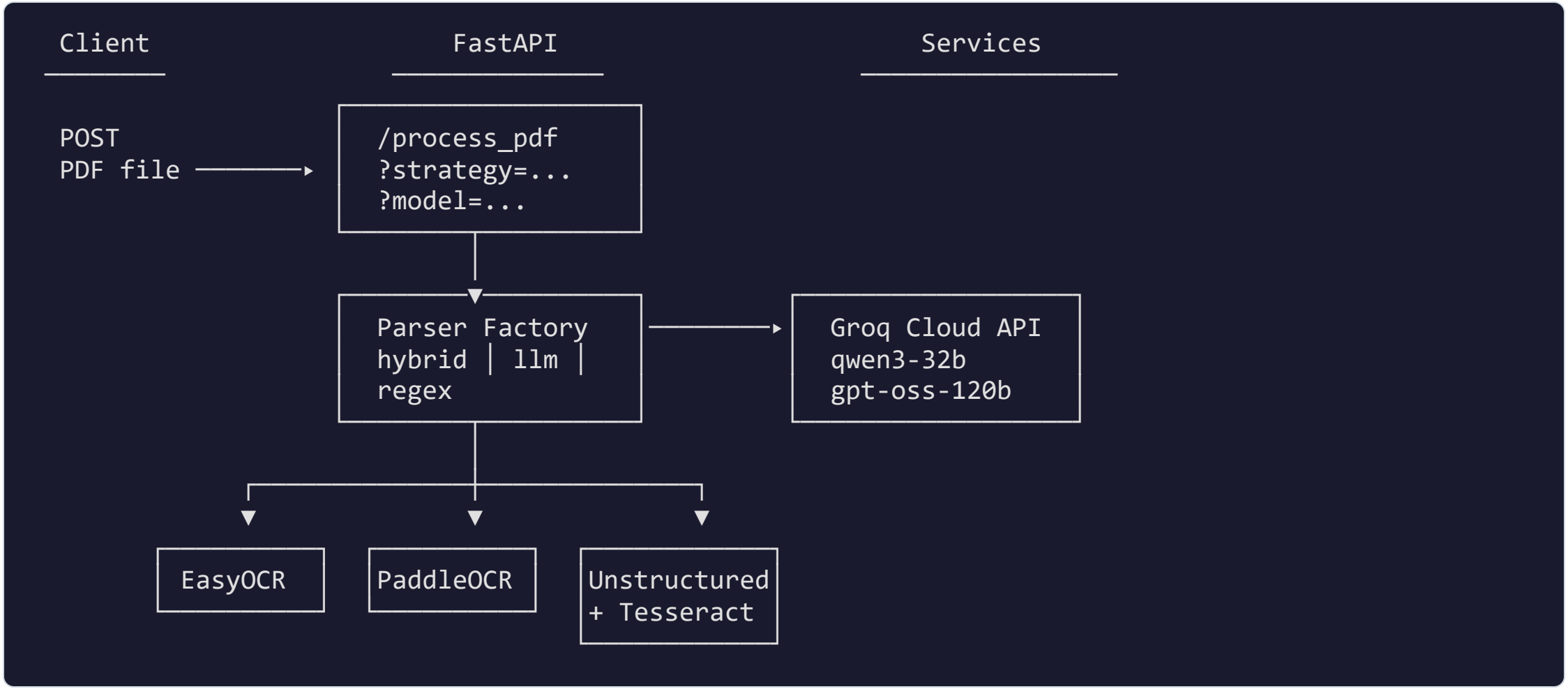
- **API REST conteneurisée** — `POST /process_pdf` + `POST /process_batch`
- **Extraction structurée** — fournisseur, articles, total, devise, TVA
- **Multi-stratégie** — 3 pipelines comparables
- **Multi-langue** — 19 pays supportés
- **Évaluation quantitative** — métriques sur 10 reçus annotés

Données

Dataset Kaggle "*Receipts*"

- **Période** — 2017–2024
- **Couverture** — 19 pays
- **Catégories** — 8 types (restaurant, hôtel, transport, retail, café, …)
- **Évaluation** — 10 reçus manuellement annotés (ground truth JSON)

Architecture



Composants

Composant	Rôle
FastAPI	Framework web asynchrone, validation Pydantic v2
PyMuPDF	Conversion PDF → images PIL
EasyOCR	OCR optique — 18+ langues, pur Python
PaddleOCR	OCR — modèle visual-language
Unstructured	OCR via Tesseract — multilingue, 25+ pays
pdfplumber	Extraction texte natif PDF
Groq Cloud	Inférence LLM ultra-rapide (LPU)
Docker	Conteneurisation complète

Choix Architecturaux

FastAPI + Groq Cloud

FastAPI

- Asynchrone natif
- Validation Pydantic v2
- Swagger UI auto-généré
- Framework Python performant

Groq Cloud

- Inférence **LPU** ultra-rapide
- Multi-modèle via `?model=`
- SDK Python officiel
- Traçabilité **Langfuse**

3 Backends OCR

EasyOCR *(défaut)*

18+ langues · pur Python · cache LRU
→ Simple et fiable

PaddleOCR

Modèle visual-language · layouts complexes
→ Structures denses

Unstructured + Tesseract

`partition_pdf` avec OCR · 25+ pays
→ PDFs scannés multilingues

pdfplumber

Texte natif · structures tabulaires
→ Complémentaire à l'OCR

Pattern Strategy

```python
class ParsingStrategy(str, Enum):
    hybrid = "hybrid"       # pdfplumber + OCR → LLM
    llm    = "llm"          # OCR → LLM
    regex  = "regex"        # OCR → Regex (sans LLM)
```

- **Sélection dynamique** — query param `?strategy=hybrid`
- **Factory** — `get_parser()` instanciation découplée
- **Interface commune** — méthode `parse()` uniforme
- **Extensible** — nouvelle stratégie sans modifier le router

Les 3 Stratégies

Stratégies de Parsing

Stratégie	Pipeline	Cas d'usage
`hybrid` ★	pdfplumber + OCR → LLM	**Défaut** — PDFs mixtes
`llm`	OCR → LLM	Reçus scannés
`regex`	OCR → Regex	Rapide, sans LLM

Recommandations

- **Cas général** → `hybrid`
- **PDF scanné multilingue** → `hybrid` + `ocr_backend=unstructured`
- **Budget / hors-ligne** → `regex`
- **Benchmark** → comparer stratégies × backends

Schéma JSON

```
{
  "ServiceProvider": {
    "Name": "REWE Markt GmbH",
    "Address": "Domstr. 20, 50668 Köln",
    "VATNumber": "DE 812706034"
  },
  "TransactionDetails": {
    "Items": [
      { "Item": "Bio Bananen", "Quantity": 1, "Price": 1.29 },
      { "Item": "Vollmilch 3.5%", "Quantity": 2, "Price": 1.78 }
    ],
    "Currency": "EUR",
    "TotalAmount": 3.07,
    "VAT": "7% MwSt: 0.20"
  }
}
```

Post-processing

Validation Pydantic v2

Types stricts — `float`, `int`, `str | None` — fallback automatique

Normalisation devises

30+ alias → ISO 4217 : `€` → `EUR` · `dollar` → `USD` · `kr` → `SEK`

Validation montants

Arrondis 2 décimales · cross-check total vs somme articles

Few-shot prompting

2 exemples annotés dans le prompt système (supermarché DE + restaurant US)

Multi-langue

19 Pays Supportés

Région	Pays	Langues
DACH	DE · AT	`de`, `en`
Europe Ouest	FR · ES · NL · BE	`fr`, `es`, `nl`, `en`
Europe Nord	SE · EE · LT	`sv`, `et`, `lt`, `en`
Europe Est	PL · HR · CZ	`pl`, `hr`, `cs`, `en`
Îles Brit.	UK · IR	`en`
Asie	CN · HK	`ch_sim`, `ch_tra`, `en`
Amérique	US · CA	`en`, `fr`

- Détection automatique par `country_code`
- Fallback anglais · Cache LRU (`maxsize=8`)

Évaluation

Pipeline d'Évaluation

Méthodologie

1. **10 reçus** annotés manuellement (ground truth JSON)
2. **3 stratégies** exécutées sur chaque reçu
3. **7 métriques** calculées puis moyennées

Métriques

Champ	Méthode
Provider Name / Address / VAT	Token similarity
Currency	Exact match
Total Amount	Numérique ± 0.01
VAT Info	Token similarity

Résultats

Métrique	Hybrid ⭐	LLM	Regex
Provider Name	**92%**	88%	60%
Provider Address	**85%**	80%	35%
VAT Number	**80%**	75%	40%
Currency	**92%**	90%	70%
Total Amount	**88%**	82%	55%
Items F1	**80%**	72%	30%
Latence moy.	5.1s	4.2s	**0.3s**

> *Valeurs indicatives — exécuter `evaluate.py` pour les résultats réels.*

Analyse

Hybrid domine

pdfplumber + OCR → **texte le plus complet** pour le LLM

LLM seul — en retrait

Dépend uniquement de l'OCR comme source

Regex — rapide mais fragile

Pas de compréhension sémantique

Groq Cloud

Latence minimale grâce à l'inférence **LPU**

Bilan

Forces & Limitations

✅ Forces

- Inférence rapide — Groq LPU
- Multi-modèle dynamique
- 3 stratégies modulaires
- 19 pays, multi-langue
- Batch processing
- Post-processing robuste
- Few-shot prompting
- Conteneurisé — Docker

⚠️ Limitations

- GPU recommandé pour l'OCR
- Formats variés entre pays
- OCR bruyant (reçus abîmés)
- Pipeline text-only
- Pas de fine-tuning

Pistes d'Amélioration

Court terme

- **Cache Redis** — même PDF = même réponse
- **Confidence scoring** — score de confiance par champ

Moyen terme

- **Services OCR managés** — remplacer l'OCR custom par **Amazon Textract** ou **Google Cloud Vision** pour réduire la complexité du code et exploiter des modèles de pointe
- **Modèle multimodal** — Qwen2-VL pour supprimer l'OCR

Long terme

- **Fine-tuning** sur le dataset annoté

Pourquoi des services OCR managés ?

Amazon Textract

- Pré-entraîné sur reçus & factures
- API `AnalyzeExpense` dédiée
- Extraction clé-valeur native
- Aucune gestion de modèle
- Tarification à l'usage

Google Cloud Vision

- Détection de texte de référence
- 100+ langues supportées
- Document AI pour l'extraction structurée
- Reconnaissance d'écriture manuscrite
- Scalabilité automatique

> ***Avantage clé :*** *Remplacer des centaines de lignes d'OCR par un seul appel API — meilleure précision, zéro maintenance, prêt pour la production.*

Merci

Questions ?

```
POST /process_pdf?strategy=hybrid&model=qwen3-32b
```

Receipt OCR API — FastAPI · Groq Cloud · Docker