**Architecture complète du projet d'extraction de données de factures et calendrier détaillé**

**I. Architecture complète du projet**

Cette architecture vise à construire un système robuste et flexible pour l'extraction automatique de données à partir de factures multi-pages et multi-formats, en intégrant les technologies de pointe de l'IA.

**0. Collecte et annotation des données**

* **Description :** Étape fondamentale pour constituer le corpus nécessaire à l'entraînement, l'évaluation et la validation des modèles.
* **Actions :**
  + Rassemblement de factures réelles (avec les accords nécessaires) ou utilisation de datasets publics pertinents.
  + Anonymisation des données sensibles pour assurer la conformité et la confidentialité.
  + Annotation manuelle précise d'un sous-ensemble des factures (identification des zones clés, des champs à extraire) pour le training supervisé.
  + Organisation logique des données en ensembles d'entraînement, de validation et de test.

**1. Acquisition et prétraitement des documents**

* **Description :** Préparation des documents sources pour l'analyse par les modèles.
* **Actions :**
  + **Décomposition PDF :** Extraction de chaque page d'un PDF multi-pages en images individuelles (format TIFF, PNG, JPEG).
  + **Nettoyage d'image :** Amélioration du contraste, suppression du bruit, redressement automatique des images scannées.
  + **Correction d'orientation :** Ajustement de l'orientation du texte si la page est scannée de travers.
  + **Segmentation automatique :** Découpage des zones pertinentes ou identification des régions d'intérêt via des modèles comme SAM (Segment Anything Model) si nécessaire.

**2. Étape OCR avancée**

* **Description :** Conversion des images en texte brut lisible par machine, en conservant la localisation spatiale.
* **Actions :**
  + Intégration d'un moteur OCR performant (ex: Tesseract fine-tuné pour les factures, ou une API OCR commerciale comme Google Cloud Vision API, Microsoft Azure Computer Vision).
  + Extraction du texte brut pour chaque mot ou bloc, associé à ses coordonnées spatiales (bounding boxes).

**3. Analyse visuelle et linguistique par VLM (Visual Language Model)**

* **Description :** Compréhension conjointe de l'apparence visuelle et du contenu textuel de la facture.
* **Actions :**
  + Le VLM prend en entrée l'image de la facture et les informations textuelles issues de l'OCR (texte + positions).
  + Il analyse la mise en page, identifie les structures (tableaux, listes), comprend les relations spatiales entre les éléments.
  + Objectif : identifier les zones clés (entêtes, pieds de page, totaux, lignes d'articles, blocs d'adresse, numéros de TVA, etc.).

**4. Compréhension contextuelle par LLM (Large Language Model) + MLM (Masked Language Model)**

* **Description :** Interprétation sémantique et amélioration de la qualité du texte extrait.
* **Actions :**
  + **LLM :** Traite le texte pour interpréter son sens, différencier les champs similaires (ex: "Date de commande" vs "Date de livraison" vs "Date de facture"), corriger des erreurs sémantiques. Il peut aussi extraire des informations non structurées (remarques, conditions).
  + **MLM :** Appliqué pour améliorer la robustesse en complétant les mots ou phrases masqués ou incomplets (typiquement des erreurs ou omissions de l'OCR) et en affinant la reconnaissance du texte.

**5. Spécialisation par Mixture of Experts (MoE)**

* **Description :** Optimisation de la performance en adaptant le traitement à la spécificité de chaque facture.
* **Actions :**
  + Le système intègre un MoE qui, grâce à un "routeur", sélectionne dynamiquement un ou plusieurs "experts" (sous-modèles spécialisés).
  + Ces experts peuvent être spécialisés par : type de facture (simple, complexe, multi-pages), langue, fournisseur spécifique, secteur d'activité, ou même par la présence de certains éléments visuels.
  + Cette approche améliore la précision en permettant à des modèles optimisés de traiter des cas précis, et la flexibilité pour s'adapter à de nouveaux formats.

**6. Structuration des données avec Structured Language Model (SLM)**

* **Description :** Transformation des informations extraites en un format exploitable et standardisé.
* **Actions :**
  + Le SLM organise les données sous forme structurée (JSON, XML, CSV).
  + Il crée des paires clé-valeur claires (ex: "Numéro de facture": "INV-2023-001").
  + Il normalise les tableaux (lignes d'articles) et consolide les informations réparties sur plusieurs pages.

**7. Validation et post-traitement**

* **Description :** Assurance qualité des données extraites et boucle d'amélioration continue.
* **Actions :**
  + **Règles métier :** Application de vérifications de cohérence (ex: le montant total correspond à la somme des lignes d'articles, validité des dates).
  + **Interface utilisateur :** Création d'une interface pour la revue manuelle des données, la correction d'éventuelles erreurs, et la validation finale (workflow semi-automatique).
  + **Apprentissage continu :** Les corrections manuelles sont utilisées pour ré-entraîner et améliorer les modèles (fine-tuning) et enrichir la base de connaissances.

**8. Intégration et export**

* **Description :** Connexion du système d'extraction avec les autres systèmes d'information de l'entreprise.
* **Actions :**
  + Développement d'APIs et de connecteurs pour exporter les données structurées vers des ERP (Enterprise Resource Planning), des logiciels comptables, des systèmes de gestion documentaire, ou des bases de données.
  + Mise en place d'outils de monitoring et de reporting pour suivre la performance du système, les taux de succès d'extraction et les erreurs résiduelles.

**II. Calendrier détaillé du projet sur 12 semaines (3 mois)**

Ce planning est conçu pour un débutant en autoformation, intégrant des phases de découverte et d'apprentissage.

|  |  |  |  |
| --- | --- | --- | --- |
| Semaines | Thème Principal | Objectifs / Tâches clés | Livrables / Résultats attendus |
| **1-2** | **Phase de préparation & Collecte de données** | - **Formation accélérée :** Fondamentaux de l'OCR, ML, LLM, VLM. Découverte des bibliothèques Python (OpenCV, Tesseract, Transformers, PyPDF2, etc.).  - **Collecte du dataset :** Identification, rassemblement et anonymisation des factures (PDF multi-pages, images).  - **Annotation :** Début de l'annotation manuelle d'un sous-ensemble représentatif des factures (champs clés, structure). | - Connaissance des concepts de base et outils.  - Corpus de factures initial prêt à l'emploi (anonymisé).  - Dataset annoté pour le training initial. |
| **3** | **Acquisition & Prétraitement (PDF -> Image)** | - Implémentation du script d'extraction des pages PDF en images.  - Intégration des fonctions de nettoyage et correction d'image (rotation, contraste).  - Tests initiaux avec SAM pour la segmentation (optionnel). | - Module fonctionnel d'extraction et prétraitement d'images depuis PDF.  - Premières images prêtes pour l'OCR. |
| **4** | **Mise en place de l'OCR Avancé** | - Intégration d'un moteur OCR (ex: pytesseract avec configuration optimisée ou API OCR).  - Extraction du texte et de ses coordonnées (bounding boxes).  - Évaluation de la qualité de l'OCR sur votre dataset (taux d'erreur). | - Module OCR fonctionnel produisant texte et coordonnées.  - Rapport d'évaluation de la performance OCR brute. |
| **5-6** | **Analyse Visuelle et Linguistique (VLM)** | - **Recherche & Sélection :** Étude et choix d'un VLM open source pertinent (ex: BLIP, LLaVA, Fuyu-8B, Qwen-VL).  - **Intégration :** Adaptation et fine-tuning du VLM pour comprendre la mise en page des factures et extraire les informations visuelles/textuelles combinées.  - **Tests :** Premiers tests sur l'identification des zones clés. | - Module VLM intégré, capable d'analyser image + texte.  - Premières extractions de zones clés (ex: entête, tableau). |
| **7** | **Compréhension Contextuelle (LLM + MLM)** | - **Intégration LLM :** Connexion à un LLM (via API ou modèle open source) pour l'interprétation contextuelle des données brutes.  - **Intégration MLM :** Utilisation d'un modèle de type MLM pour la complétion et correction des erreurs d'OCR et de texte.  - **Tests :** Évaluation de l'amélioration de la précision grâce à la correction sémantique. | - Module LLM/MLM capable d'affiner le texte extrait et de corriger les erreurs.  - Données plus propres et sémantiquement cohérentes. |
| **8** | **Spécialisation par MoE (MVP)** | - **Compréhension MoE :** Approfondissement du concept de Mixture of Experts.  - **Implémentation simplifiée :** Mise en place d'un routage simple basé sur des règles (ex: détection de la langue, présence de mots clés pour un type de facture).  - **Spécialisation :** Création de 2-3 "experts" rudimentaires (ex: un expert pour factures simples, un pour factures avec tableaux). | - Ébauche d'un système MoE avec un routage basé sur des règles simples.  - Capacité à différencier et appliquer un traitement légèrement différent selon le type de facture. |
| **9** | **Structuration des données (SLM)** | - Développement du module de transformation des données extraites en format structuré (JSON ou CSV).  - Gestion de l'aggrégation et de la normalisation des données multi-pages (ex: lignes d'articles réparties sur 2 pages). | - Module SLM fonctionnel produisant des données structurées et consolidées.  - Fichiers JSON/CSV contenant les informations clés de la facture. |
| **10** | **Validation & Interface utilisateur (MVP)** | - Implémentation des règles de validation métier basiques (ex: montant\_total = somme(lignes\_HT) + TVA).  - Création d'une interface web/CLI simple pour la revue manuelle et la correction des données (ex: affichage des champs et validation par l'utilisateur). | - Module de validation automatique.  - Interface utilisateur rudimentaire pour la correction manuelle.  - Workflow semi-automatique pour le traitement des factures. |
| **11** | **Intégration & Export** | - Développement d'un connecteur ou d'une API simple pour exporter les données structurées (JSON/CSV) vers une base de données locale ou un fichier.  - Mise en place d'une logique de monitoring basique (nombre de factures traitées, taux de succès). | - Fonctionnalité d'export des données.  - Logique de monitoring simple. |
| **12** | **Tests finaux, Optimisation & Documentation** | - **Tests complets :** Tester le pipeline de bout en bout sur un jeu de données de test varié.  - **Optimisation :** Identifier les goulots d'étranglement, améliorer les performances et la précision.  - **Documentation :** Rédiger la documentation technique du projet et un rapport final (incluant état de l'art et état des lieux). | - Système fonctionnel et testé.  - Rapports de performance.  - Documentation technique complète.  - Rapport de projet final. |