

Extraire les infos d'un lot de documents

Un tableau traçable - et la différence entre trouvé, non trouvé et illisible

Module DO1 - FAIRE - 7 min - version 1.1

À vérifier à l'écran avant de suivre ce module

- Nombre de fichiers simultanés : [À VÉRIFIER : combien de documents peuvent être joints à une même conversation, par offre]
- Taille et formats : [À VÉRIFIER : taille maximale par fichier et formats acceptés]
- PDF scannés : [À VÉRIFIER : les PDF images (scans) sont-ils lus, et avec quelle fiabilité ?]

Volatilité V3 - aucune de ces données produit n'est déduite. Relevé réel requis avant publication.

Prix, limites et noms de menus changent sans prévenir. Si ce que tu vois à l'écran ne correspond pas, c'est l'écran qui a raison.

Tu dois retrouver la même information dans plusieurs documents. L'IA peut aider à faire une première extraction, mais le tableau final doit rester traçable jusqu'au document source.

Ce que tu sauras faire en sortant

- Poser une question d'extraction unique à plusieurs documents.
- Exiger document + citation exacte pour chaque résultat retenu.
- Ne jamais confondre « non trouvé » avec « absent du document ».
- Effectuer un contrôle positif avant toute conclusion d'absence.

L'essentiel, en trente secondes

Quatre états possibles :

- TROUVÉ - une réponse + une citation exacte localisable.
- NON TROUVÉ - la recherche n'a pas permis de retrouver l'information ; aucune conclusion d'absence.
- ILLISIBLE / NON EXPLOITABLE - le document n'a pas pu être lu de façon suffisante.
- ABSENCE CONFIRMÉE - uniquement après un contrôle positif dans le document avec une méthode adaptée.

Le quatrième état est volontairement plus exigeant. Une absence ne se déduit pas de l'échec du modèle à trouver.

Le prompt d'extraction

``text Pour chaque document, cherche uniquement : [QUESTION]. Rends : document | état | réponse | citation exacte | repère dans le document.

États autorisés : - TROUVÉ : citation exacte disponible ; - NON TROUVÉ : tu n'as pas retrouvé l'information ; - ILLISIBLE : tu ne peux pas lire suffisamment le document.

N'écris jamais « absent » : cette conclusion sera vérifiée séparément. N'invente ni réponse ni citation. ``

Le tableau n'est pas la preuve. La citation ouverte dans le document est la preuve du résultat « trouvé ».

Micro-défi

Le tableau indique NON TROUVÉ pour un document. Quelle phrase peux-tu écrire ?

- « Le document ne contient pas cette information. »
- « L'information n'a pas été retrouvée pendant cette extraction. »
- « L'information est probablement absente. »

Attendu : la 2. Pour passer à « absent », il faut un contrôle positif séparé.

Confirmer une absence : le contrôle positif

Pour un document où l'information compte vraiment :

- ouvre le document ;
- cherche un ou plusieurs termes attendus ou synonymes distinctifs ;
- vérifie les sections où l'information devrait logiquement apparaître ;
- si le document est une image/scan ou si la recherche textuelle est impossible, utilise une méthode de lecture adaptée au document ;
- note ce qui a été cherché, où, et avec quel résultat.

Seulement après ce contrôle peux-tu écrire, avec la portée exacte :

« Information non retrouvée après contrôle dans [sections/termes vérifiés] le [date]. »

Même cette formulation est préférable à une affirmation absolue « le document ne contient pas X » lorsque le document ou la méthode de recherche comporte des limites.

Contrôle des citations

Pour chaque ligne importante, retrouve la citation dans l'original. Si tu fais un échantillonnage pour gagner du temps, il ne certifie que les lignes contrôlées.

Le module ne fournit plus de nombre magique du type « trois citations suffisent » ou « cinq minutes ».

La profondeur de contrôle dépend de l'impact :

- faible impact contrôle par échantillon documenté possible ;
- décision engageante vérification des lignes utilisées dans la décision ;
- citation introuvable la ligne concernée est rejetée et le processus d'extraction est réévalué.

Une citation fausse ne prouve pas automatiquement que toutes les autres sont fausses, mais elle retire le droit de traiter le lot comme fiable sans contrôles supplémentaires.

Le tri F3 avant le lot

Chaque document passe le tri F3 avant envoi à un service tiers.

Pour l'exercice, utilise des documents fictifs, publics appropriés ou correctement préparés. Si la matière réelle contient des données de tiers, des secrets ou un cadre contractuel particulier, le module ne donne pas l'autorisation de la téléverser.

La recommandation « recopier une section » n'est pas automatiquement suffisante : la section recopiée doit elle aussi passer le tri F3.

Le cas où ça se passe mal

Cas reconstitué.

Un lot contient des documents texte et des scans. La sortie remplit toutes les lignes, y compris celles provenant des scans, mais aucune citation n'est localisable pour certains fichiers.

Correction : les lignes sans citation passent en NON TROUVÉ ou ILLISIBLE. Tu ne transformes pas une ligne remplie en preuve parce que le tableau est complet.

Ce que ce module ne règle pas

- Il ne qualifie pas juridiquement une clause.
- Il ne prouve pas qu'une information est absente à partir d'un échec de recherche.
- Il ne garantit pas la lecture des scans : c'est un marqueur produit ouvert.
- Il ne donne aucun seuil universel de taille d'échantillon.
- Il ne remplace pas un système documentaire adapté à de très grands volumes.

Quand tu t'arrêtes

- Le document n'a pas passé le tri F3.
- Une information engageante n'a pas de citation localisable.
- Le document est illisible avec la méthode disponible.
- Tu es sur le point de transformer NON TROUVÉ en ABSENT sans contrôle positif.

À toi

Crée cinq mini-documents fictifs. Dans trois, place une information précise ; dans un quatrième, utilise un synonyme ; dans le cinquième, ne mets volontairement rien.

- Lance l'extraction avec les quatre états.
- Vérifie chaque citation TROUVÉ dans l'original.
- Pour le document sans résultat, fais le contrôle positif et documente les termes/sections vérifiés.
- Compare NON TROUVÉ avant contrôle et la conclusion bornée après contrôle.

C'est réussi si tu n'utilises jamais « absent » comme sortie automatique du modèle.

Vérifie ton geste

- Pas de citation. Pas de résultat retenu comme TROUVÉ.
- Le modèle écrit "absent". Reformule en NON TROUVÉ puis contrôle séparément.
- PDF image illisible. ILLISIBLE, pas absence.
- Citation retrouvée. Cela prouve le passage cité, pas ton interprétation juridique.
- Une citation du lot est fausse. Rejette la ligne et augmente le contrôle avant d'utiliser le reste.

Seuil de réussite : 4 sur 5, questions 2 et 3 obligatoires.

Selon ce que tu obtiens

- Plusieurs NON TROUVÉ normal : ils signalent où le contrôle manuel est nécessaire.
- Une ligne ILLISIBLE bon échec ; n'invente pas pour compléter le tableau.
- Une absence confirmée après contrôle note la méthode et la date, avec une formulation bornée.
- Bloqué travaille sur trois documents fictifs et une question dont tu connais la réponse dans deux d'entre eux.

Fiche mémo

- Une question d'extraction par passe.
- TROUVÉ = citation exacte localisable.
- NON TROUVÉ absent.
- ILLISIBLE absent.
- Absence contrôle positif documenté.
- Impact élevé vérifie toutes les lignes utilisées pour décider.

Ce que tu viens de pratiquer

- Tu as rendu l'extraction traçable.
- Tu as séparé absence et échec de recherche.
- Tu as appliqué le contrôle positif exigé par la gouvernance du parcours.

Sources et statut : docs/bloc-documents-PROVENANCE.md`. DO1 reste S2 et brouillon ; les trois marqueurs produit restent ouverts.