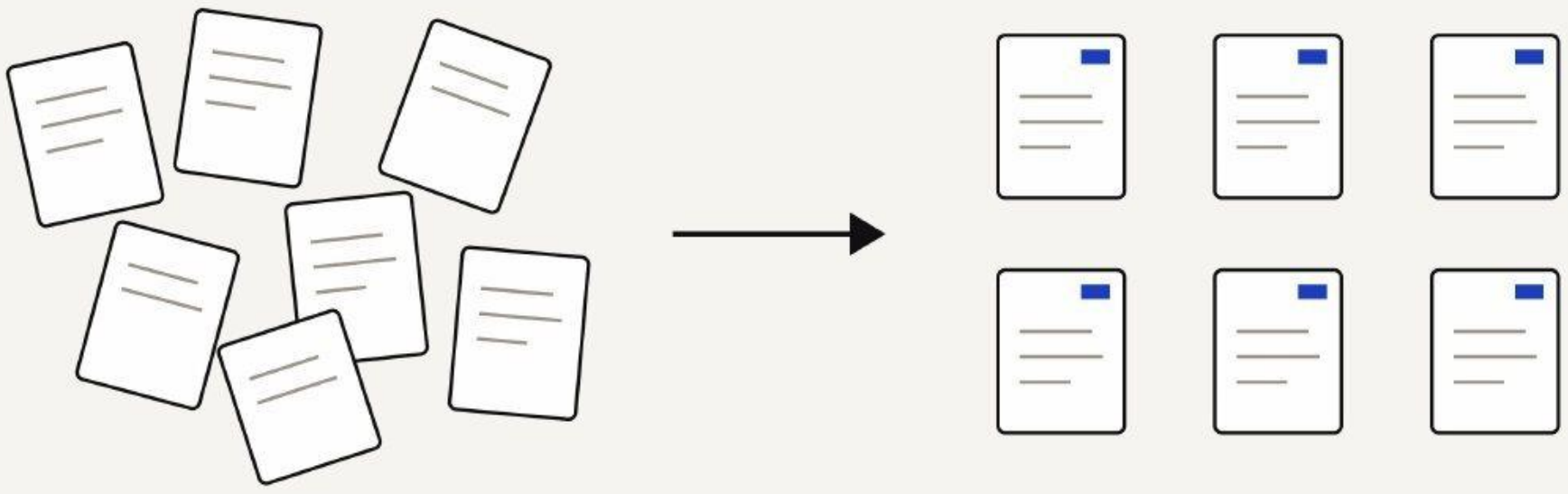
PREMIERS RÉSULTATS

« Pas d'IA sans données propres. » Et si l'IA les rendait propres ?

Rendre un fonds documentaire réel exploitable par l'IA, sans chantier de rangement préalable.

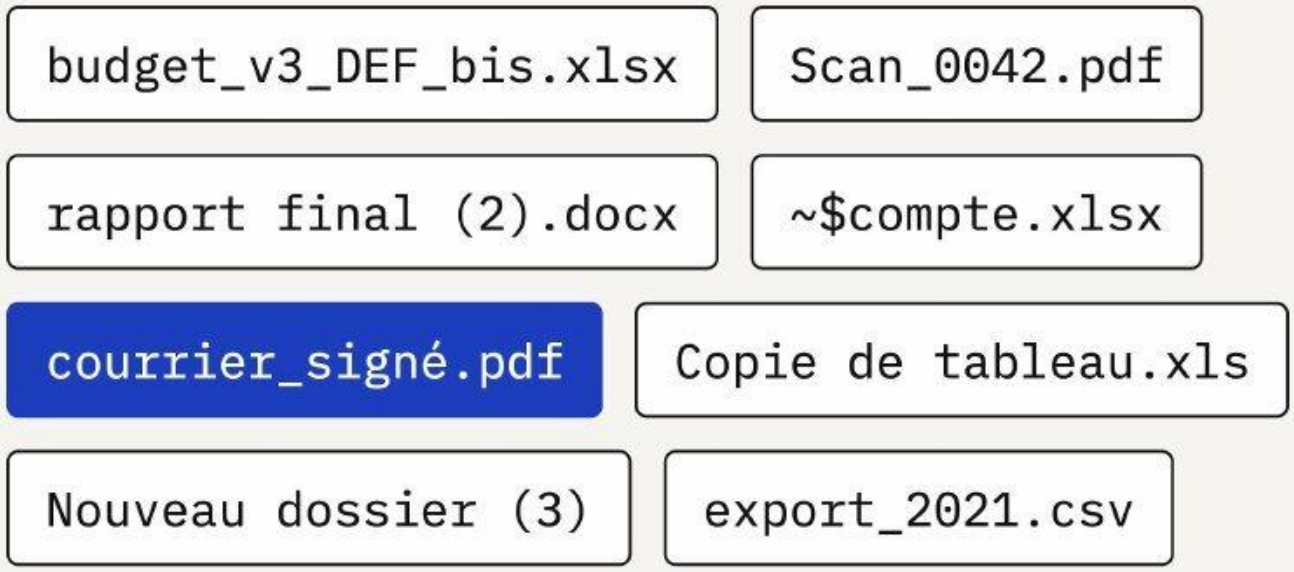


Faites défiler →

LE CONSTAT

L'IA du quotidien ne voit pas ce que l'organisation sait.

Ce qu'une organisation sait est dispersé : dossiers partagés, logiciels métiers, archives, courriels. Des documents hétérogènes, en plusieurs versions, mal nommés, souvent scannés. Les courriers qui comptent sont noyés dans le reste.



On utilise l'IA sur des questions générales, sans accès à cette mémoire. Et l'on croit qu'il faudrait d'abord tout nettoyer.

LE TEST

Un des fonds les plus difficiles qui soient.

Des documents réels sur l'organisation et le financement d'ESSMS : des chiffres partout, des procédures, des versions successives.

1 800
fichiers

6
exercices

12
formats, dont des classeurs de quatre générations

80 %
des pages PDF sans aucun texte

≈ 50 M de jetons de texte, soit plus d'une année de lecture à temps plein

Rien n'a été rangé avant de commencer.

→

RÉSULTAT 1 · DATER

De quelle année parle ce document ?

Le nom du fichier et le dossier où il est rangé ne suffisent pas à le savoir. C'est le cas dans la plupart des organisations.

Se fier au nom du fichier	**juste 1 fois sur 8**
Se fier au dossier de rangement	**faux 1 fois sur 3**
Se fier aux dates du fichier	**elles ne datent rien**

98,5 % des documents rattachés à leur année par l'IA, qui lit le contenu comme le ferait un professionnel

L'IA classe aussi chaque pièce et repère celles qui manquent.

→

RÉSULTAT 2 · IDENTIFIER

De quel établissement parle ce document ?

Un logiciel métier reconnaît un établissement à son identifiant, le numéro FINESS. Les documents, eux, le désignent par son nom, son sigle ou sa ville.

EXEMPLE FICTIF

- Maison d'enfants Les Tilleuls
- MECS « Les Tilleuls »
- la MET
- le foyer de Saint-Rémy

→ **Un seul et même établissement**

2 sur 3

des mentions d'un établissement ne portent aucun FINESS

1 sur 62

rapprochements s'appuie sur un identifiant

L'IA fait le lien comme un professionnel. L'expert valide des centaines de propositions, pas des milliers de pièces.

RÉSULTAT 3

Pour qualifier le fonds, un modèle ouvert en local suffit.

Une très bonne nouvelle pour la confidentialité.

AUSSI BIEN

70 %
de pièces aux informations clés toutes justes, contre 68 à 80 % pour les modèles commerciaux

5 sur 5
établissements bien identifiés, numéros FINESS compris

SES LIMITES

142 min
pour qualifier 40 pièces, contre 9 avec un modèle commercial

1 sur 9
lecture exploitable d'un long dossier d'un seul tenant, contre 9 sur 9

Modèle d'environ 30 milliards de paramètres, en 8 bits, sur une machine de 96 Go installable chez un client (du type DGX Spark ou Mac Studio).

Pour lire de longs dossiers, une instance cloud dédiée n'a rien de saugrenu.

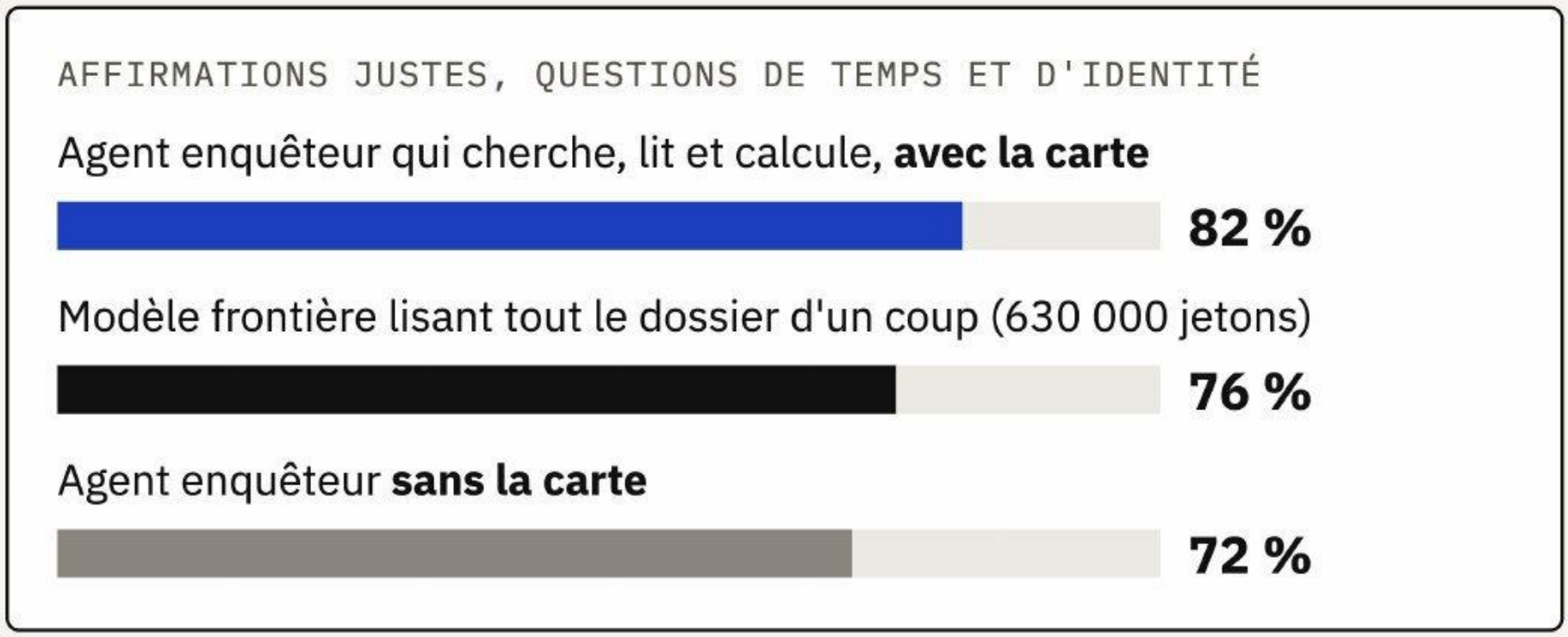
RÉSULTAT 4 · TROUVER

Trouver la bonne information : oui.

LA CARTE = ce que l'IA a établi pour chaque document : sa nature, son année, l'établissement concerné.

197 / 200
chiffres justes, chacun avec sa source

1 / 1 934
citation introuvable dans la pièce citée



Sans la carte, l'agent se trompe une fois sur deux quand les sources se contredisent.

Mémoire sophistiquée ou non : c'est la carte qui compte.

→

LA LIMITE

Analyser comme un expert : pas encore.

≈ 1 / 2

des points d'un rapport d'expert retrouvés, avant toute instruction

+ 4

points pertinents ajoutés par réponse, absents du rapport

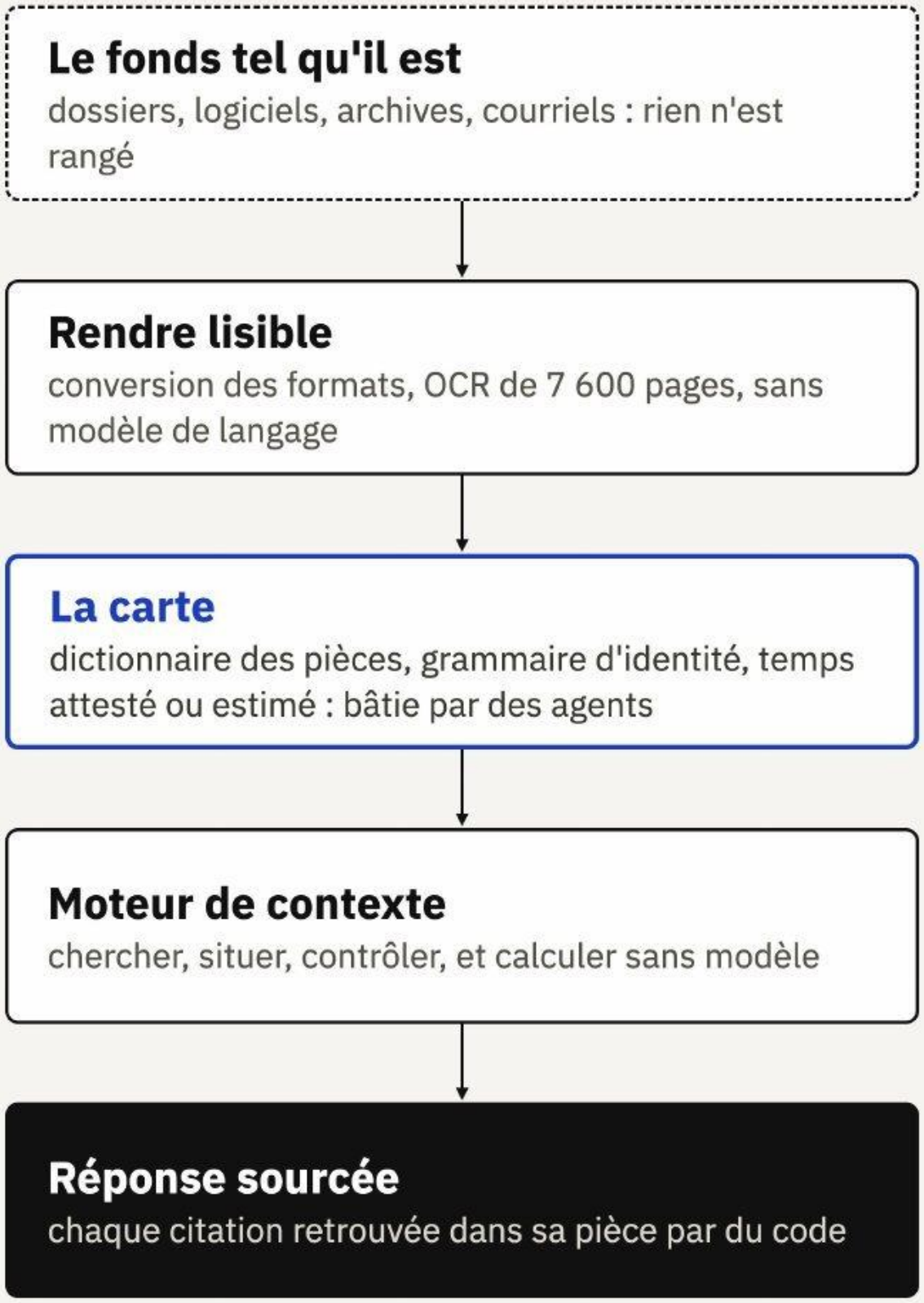
01 Elle confond encore le prévu et le réalisé, même quand la consigne le lui interdit.

02 Face à une pièce qui se contredit, 5 réponses sur 7 la suivent à la lettre. L'expert lui-même l'avait d'abord mal lue.

03 Les modèles ouverts datent mal (57 % contre environ 80 %), confondent versions et colonnes, et inventent une justification à un chiffre faux.

L'IA trouve. L'expert analyse.

Du vrac à la réponse sourcée



HORIZON DE RECHERCHE

Des traces aux actes

D · Traces
tableaux, rapports, courriers

S · Actes et positions
demander, proposer, contester, décider

G · Grammaire latente
régularités composables

$$p(S \mid D, C) \propto p_\theta(S \mid C) \cdot p_\varphi(D \mid S, C)$$

Un montant n'est pas une valeur : c'est une assertion. Qui l'avance, à quel stade, dans quelle version, connue quand ?

HYPOTHÈSE EXPLORATOIRE

Travaux voisins. Feldman & Pentland (2003), routines organisationnelles · Berti et al. (2023), OCEL 2.0, journaux d'événements centrés objets · Chambers & Jurafsky (2009), schémas narratifs · Kim, Dyer & Rush (2019), induction de grammaire · Dung (1995), argumentation · W3C PROV-O, provenance · Grünwald (2007), longueur de description minimale.

CE QUE ÇA CHANGE

01 **Pas de chantier de rangement préalable.**

02 **Un fonds réel devient interrogeable, avec ses sources.**

03 **L'IA trouve. L'expert analyse.**

Une recherche en cours, sur un premier terrain. Prochaines étapes : un second terrain, puis un corpus fictionnalisé pour une mesure indépendante.

REPERES

Structurer votre mémoire organisationnelle pour la rendre exploitable par des outils IA au quotidien.

reperes.bmse.fr