

ESTIMER L'EMPREINTE CARBONE VIA LLM : VERS UNE ARCHITECTURE FRUGALE ET ADAPTATIVE FACE AU PARADOXE ÉNERGÉTIQUE

Dr. Annia Abtout &
Dr. Natalia Kalashnikova

Centre de Recherche & Innovation Talan

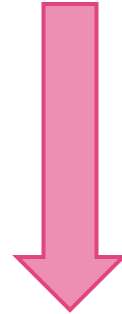
30.06.2026

PFIA ' 26

TALAN

CONTEXTE MÉTIER

Besoin métier	AO / réponses aux AO	Formulaire manuel
Calculer l'empreinte carbone des missions IT à l'étape d'avant-vente à partir des AO / réponses aux AO	• Longs • Hétérogènes • Sans format unifié • Information souvent implicite, incomplète ou absente	• Chronophage • Source d'erreurs

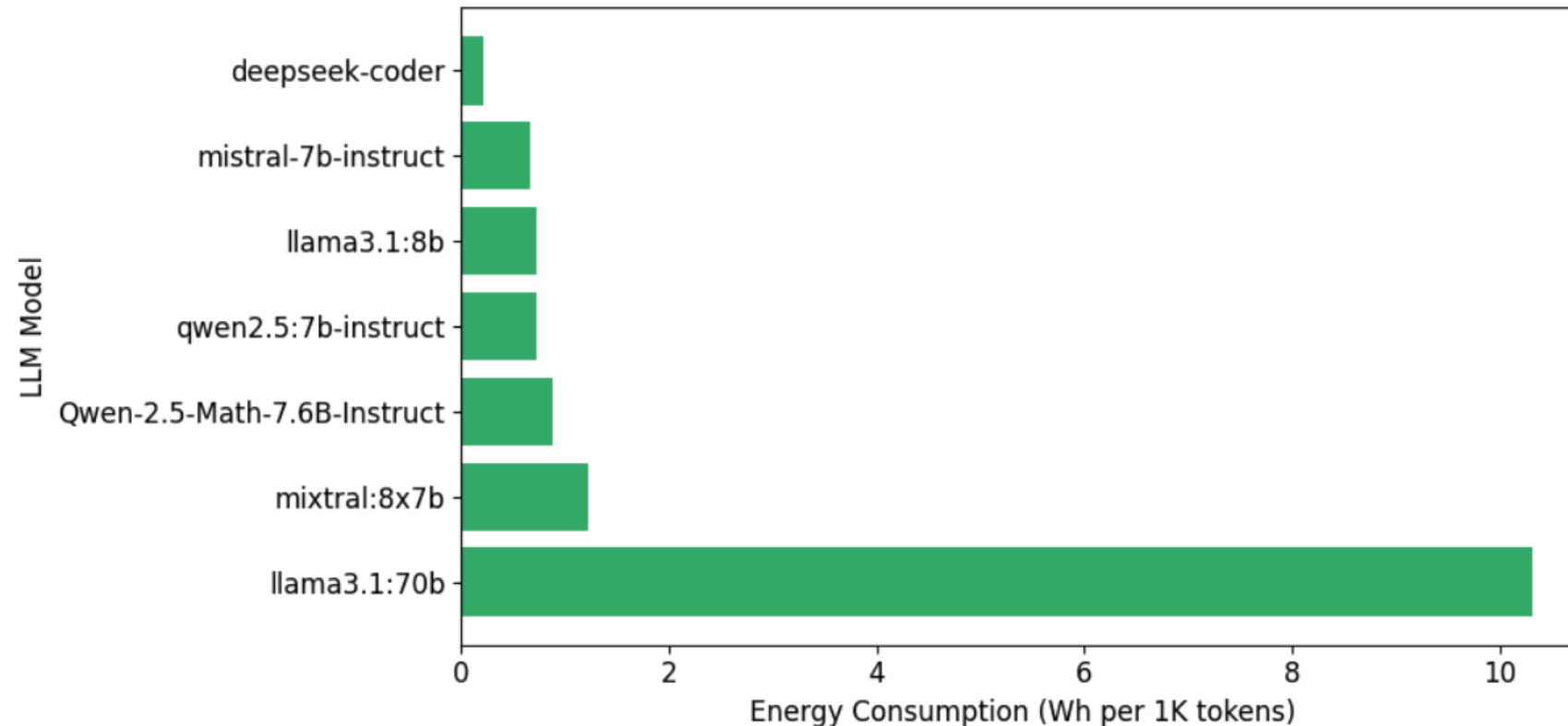


Automatisation de remplissage du formulaire avec IA Gen

- Améliorer la précision du formulaire
- Diminuer la durée de remplissage
- Prévion des données manquantes grâce au système RAG

L'IMPACT ÉNERGETIQUE DES LLM

Utiliser un LLM pour estimer l'empreinte carbone d'un AO... génère lui-même une empreinte carbone.



Kanso & Khazem, Talan, 2025 (mesure CodeCarbon)

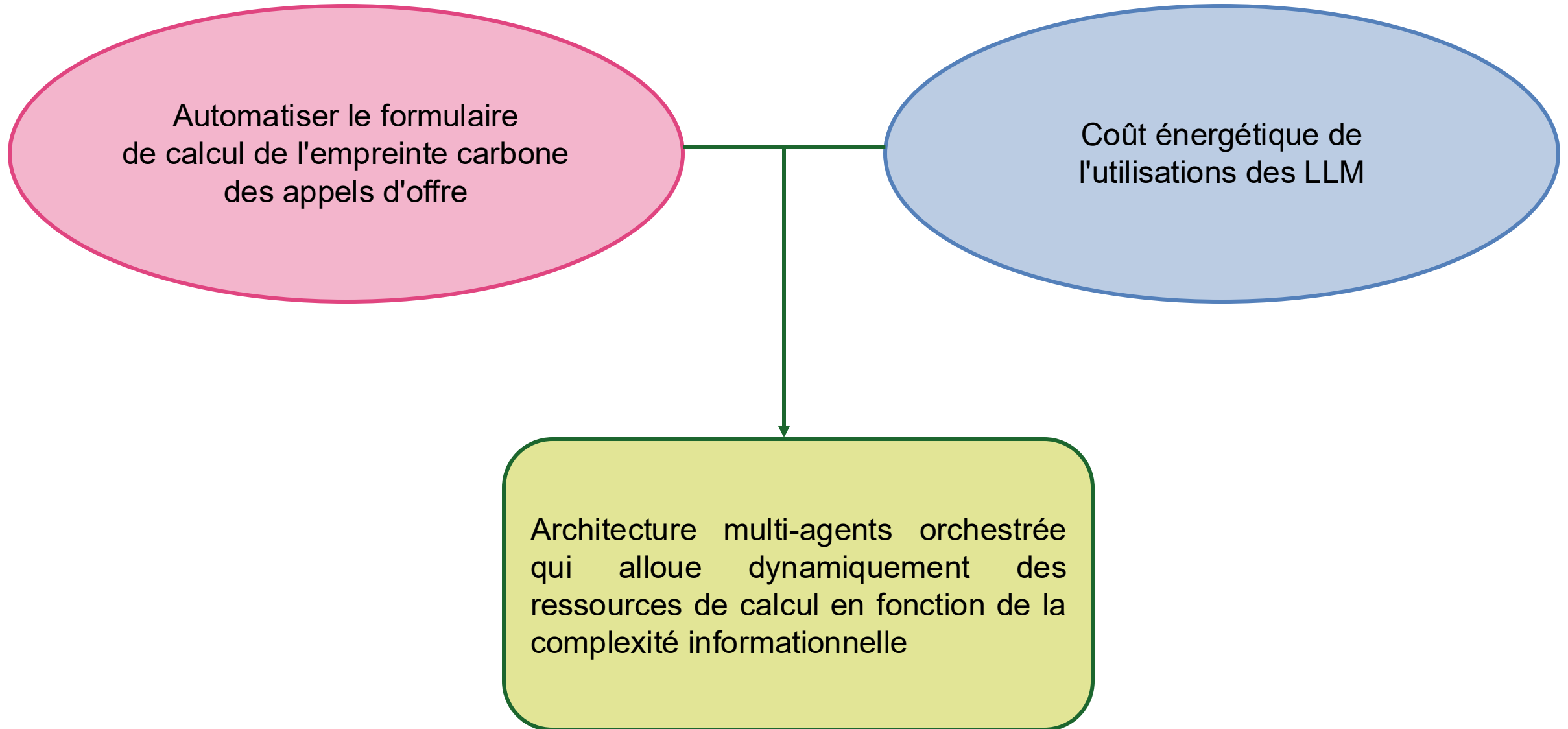
Un LLM n'est pas gratuit en énergie

Facteur $\times 50$ entre le modèle le plus frugal et le plus gourmand pour 1000 tokens dans l'étude de Kenso et al.

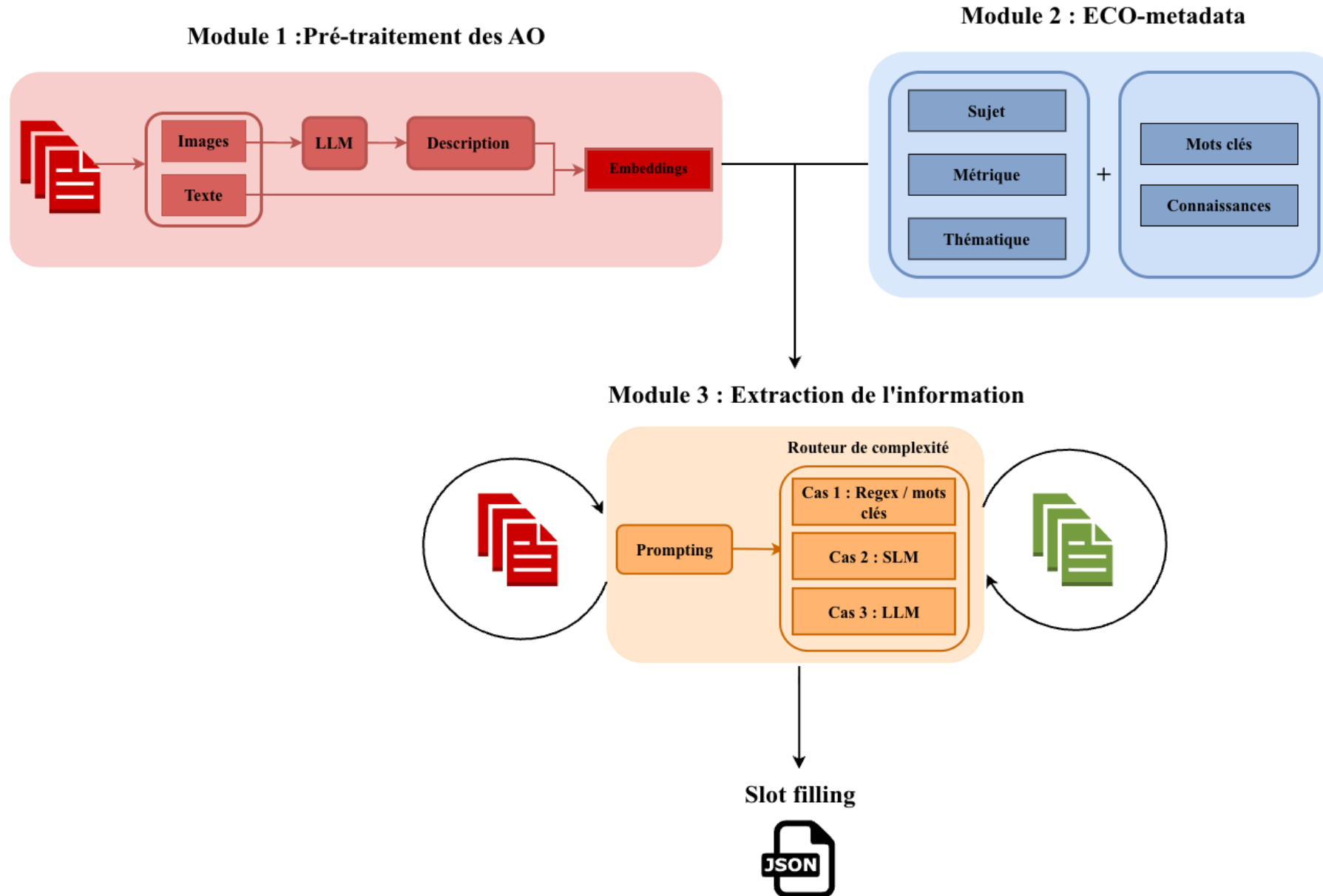
Le paradoxe

L'outil d'optimisation de l'empreinte carbone devient lui-même une source d'empreinte. Appliquer un LLM massivement sur des centaines d'AO n'est pas anodin.

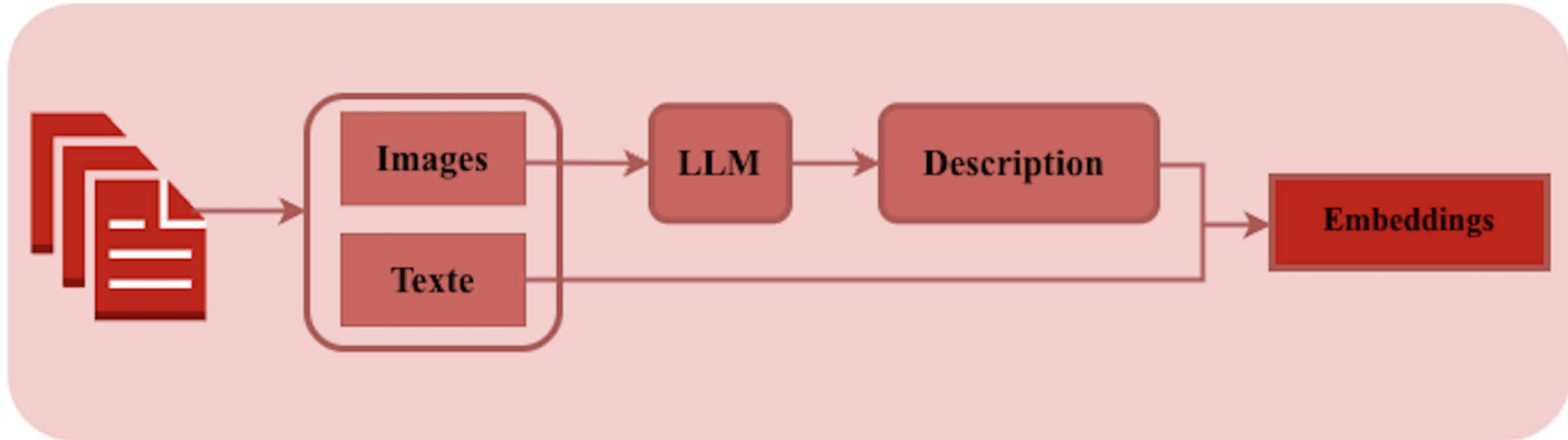
LE PARADOXE ENERGETIQUE



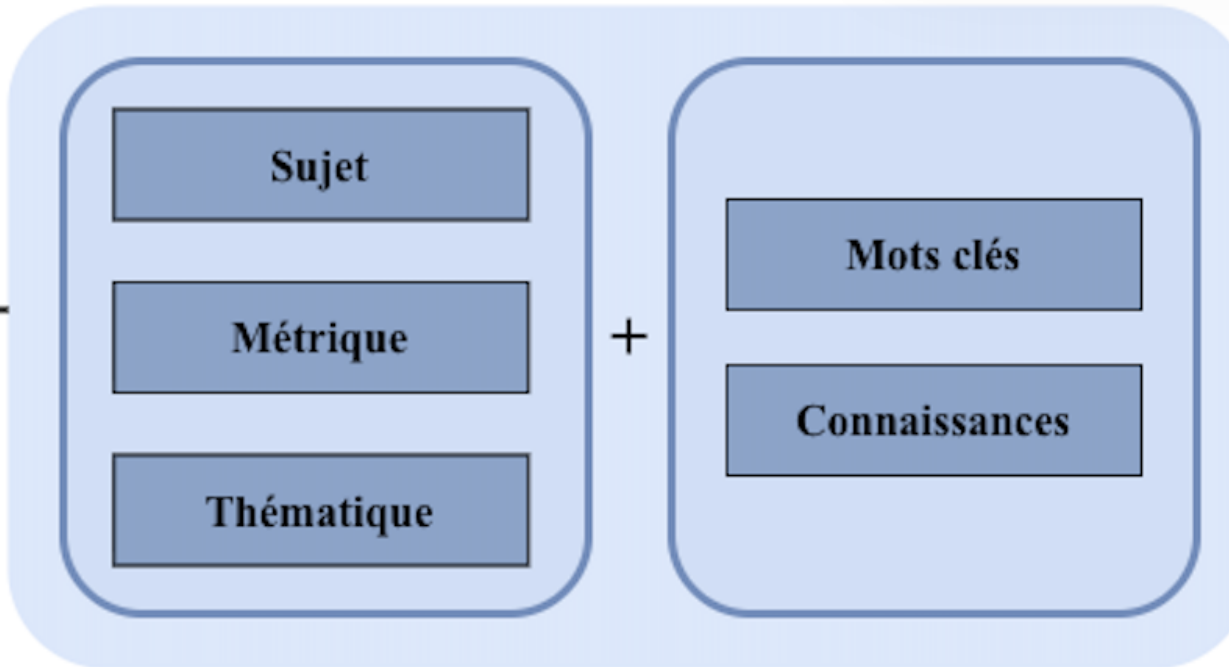
NOTRE SOLUTION : HOPE



HOPE : MODULE DE PRÉTRAITEMENT DES AO



HOPE : MODULE ECO-METADATA



Exemple de metadata pour l'extraction de champs liés aux ressources de calcul

Contexte metier

Thématique : Numerique & Infrastructure de calcul

Sujet :

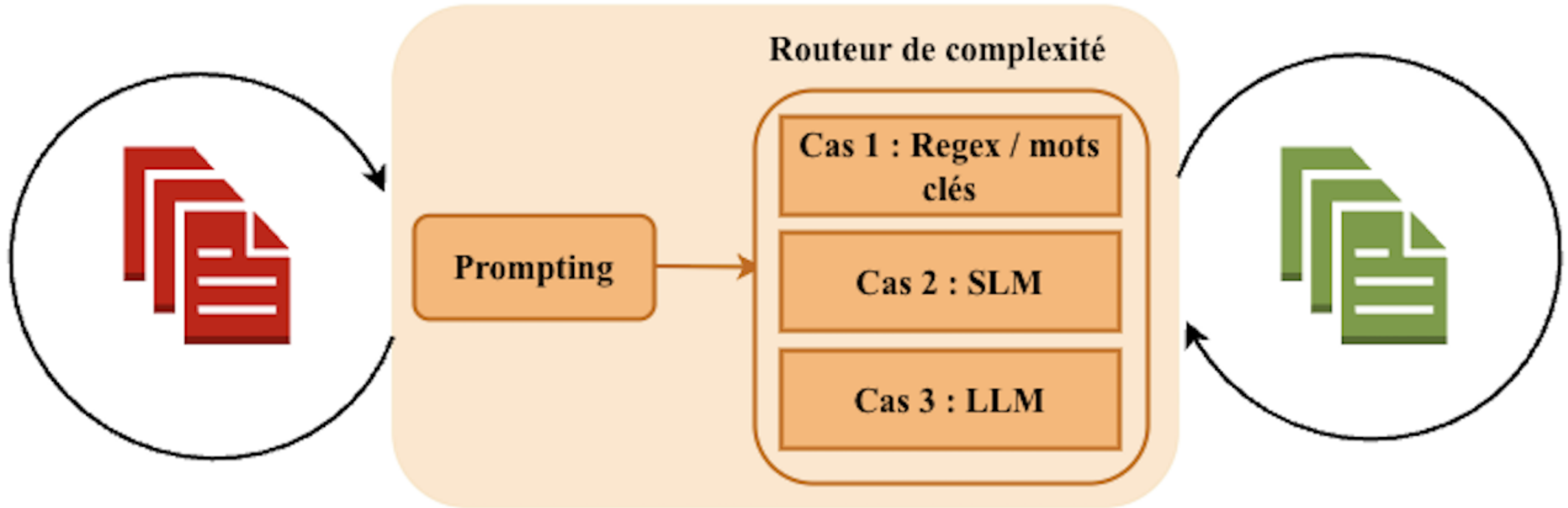
Type de ressources computationnelles mobilisees

Metrique :

Type d'infrastructure (GPU/CPU/Cloud/On-premise), fournisseur si cloud, localisation si mentionnee

Mots-cles : GPU, CPU, cloud, serveur, infrastructure, AWS, Azure, GCP, on-premise, datacenter, entraînement, inference.

HOPE : MODULE D'EXTRACTION DE L'INFORMATION



La contribution centrale de HOPE : n'escalader vers un modèle plus coûteux que si le niveau précédent a échoué.

CONFIGURATIONS ENVISAGÉES

PROMPTING

- One-shot
- Few-shot
- Chain-of-Thought

CHUNKING

- Par page
- Hiérarchique

EMBEDDINGS

- Small : text-embedding-3-small (1536), qwen3-embedding-0.6b (1024)
- Large : text-embedding-3-large (3072), qwen3-embedding-8B (4096)

INFÉRENCE

- SLM : gptO4-mini, qwen3.5-2b, mistral-small
- LLM : gpt-5.5, qwen3.5-9b, qwen3.5-27b, qwen3.5-35b, mistral-large-3

HOPE : CORPUS & ANNOTATION

100

Appels d'offres

3

Annotateurs par
document

Processus d'annotation

1. Parcourir chaque AO
2. Renseigner le formulaire carbone
3. Indiquer si info absente / déduite / fiable
4. Vote majoritaire des 3 annotations -> gold standard

Critères de construction du corpus

Diversité des domaines

Formats variés - reflet de l'activité réelle de TALAN

Variété des tailles de documents

Sensibilité au contexte long – robustesse du chunking

HOPE : PROTOCOLE D'ÉVALUATION HYBRIDE (SCORE HYBRIDE)

Champs déterministes & explicites

Exemples : dates, lieux, durées, etc.

Précision

Proportion d'extractions correctes

Rappel

Taux de champs correctement récupérés

F1-Score

Moyenne harmonique précision/rappel

Exact Match

Correspondance exacte valeur extraite / gold standard

Faible coût computationnel

Champs sémantiques & implicites

Exemples : infrastructure, mobilité, hébergement

Framework Grina & Kalashnikova (2025)

- BERTScore, ROUGE, F1 (évaluation quantitative)
- LLM-as-a-judge (évaluation qualitative)
- Détection de conflits -> validation humaine si écart > seuil

Score
quantitatif

Score
LLM-as-a-judge

Validation
humaine

Plus coûteux - réservé aux champs nécessitant une inférence sémantique

Principe de frugalité : méthodes coûteuses uniquement si nécessaire

HOPE : ÉCO-PRÉCISION - MÉTRIQUE COMPOSITE

Évaluer conjointement la qualité d'extraction et l'empreinte carbone de l'inférence.

$$\text{ÉCO-PRÉCISION} = \frac{\text{SCORE HYBRIDE}}{\text{LOG}(1 + \text{EMPREINTE CARBONE})}$$



Principe : la métrique pénalise les architectures qui consomment trop d'énergie, même si elles extraient correctement.

Une extraction parfaite avec un LLM sera moins bien notée qu'une extraction légèrement imparfaite avec un SLM.

Score hybride

Agrégation des scores d'extraction sur l'ensemble des champs du formulaire carbone

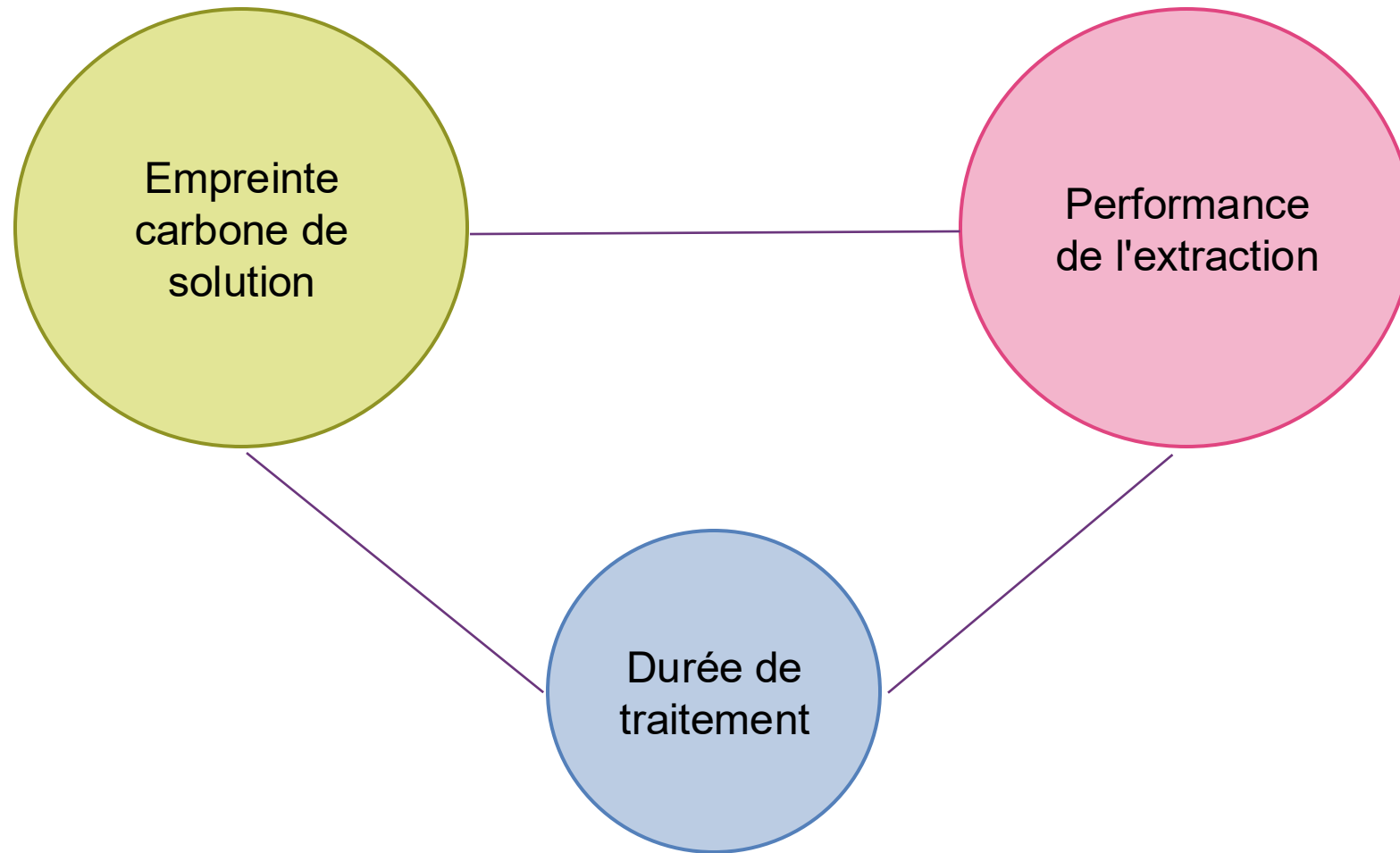
$\log(1 + \text{empreinte})$

Compression logarithmique pour couvrir les ordres de grandeur de routeur

CodeCarbon

Sonde logicielle pour mesurer l'empreinte carbone réelle de chaque appel d'inférence

TRAVAUX EN COURS ET LIMITES



Merci



CONTACTS

ABTOUT Annia
annia.abtout@talan.com

KALASHNIKOVA Natalia
natalia.kalashnikova@talan.com

PROTEUS. ÉVALUATION

