Outil NARVAL : extraction automatisée d'indicateurs SISPEA à partir de rapports PDF

Geneviève Fleury, Malou Berthe, Benoit Lefeuvre, Sophie Portela et Grégoire Etot
Direction Surveillance, Evaluation, Données (DSUED),
Office Français de la Biodiversité (OFB), Vincennes, France
(10 avril 2025)

L'observatoire national des services publics d'eau et d'assainissement centralise dans le système d'information SISPEA des données sur l'organisation des collectivités, la tarification et les performances des services publics d'eau et d'assainissement en France. Conformément à la loi NOTRe, ces données sont bancarisées chaque année sur SISPEA par les collectivités en charge de la compétence eau et/ou assainissement, au travers d'une liste d'indicateurs réglementaires codifiés. Pour faciliter cette tâche de saisie et enrichir la base de données, nous avons développé l'outil NARVAL permettant d'extraire les valeurs des indicateurs SISPEA à partir de rapports au format PDF publiés par les collectivités ou leurs délégataires. L'approche choisie se veut frugale et repose sur l'utilisation de bibliothèques python et de petits modèles de langage (SLM) ouverts : les indicateurs sont d'abord extraits de certains tableaux récapitulatifs puis, à défaut, un SLM est interrogé pour sonder le texte et les autres tableaux. A ce jour, NARVAL est limité aux indicateurs d'assainissement collectif et ne traite que les PDFs non scannés de « petites » collectivités. Plusieurs pistes d'amélioration sont discutées dans ce rapport, aussi bien pour améliorer les performances de NARVAL que pour généraliser son usage.

I. INTRODUCTION

Le besoin d'extraire des données structurées à partir de documents PDF est partagé par de nombreuses entreprises ou administrations publiques, cherchant à valoriser les données contenues dans des rapports, factures ou encore formulaires, produits en masse mais difficilement exploitables en l'état.[1–4] L'automatisation de cette tâche se heurte toutefois à des défis techniques majeurs, notamment lorsque les documents présentent des structures hétérogènes ou des mises en page complexes. En effet, les méthodes traditionnelles dites heuristiques, c'est-à-dire basées sur des règles explicites (détection de mots-clés, analyse de mise en page, extraction de tableaux, ...), montrent rapidement leurs limites dès que la structure d'un PDF s'écarte du modèle calibré. Face à cette difficulté, les avancées récentes dans le domaine de l'apprentissage profond et plus généralement de l'intelligence artificielle (IA), ont ouvert de nouvelles perspectives. En particulier, les grands modèles de langage (LLM) et les approches hybrides couplant vision par ordinateur et traitement du langage permettent aujourd'hui une interprétation plus flexible et contextuelle des PDFs, en exploitant des capacités de compréhension sémantique et de raisonnement.[5–7]

Si les approches par IA marquent un tournant en matière d'extraction automatisée de données, leur mise en œuvre soulève des problèmes importants : coûts financiers élevés, consommation excessive d'énergie et de ressources, dépendance technologique, risques de sécurité, manque de transparence, enjeux éthiques, ... Face à ces limites, émerge notamment en France et en Europe une volonté de promouvoir une IA sobre, ouverte et maîtrisée.[8–10] En particulier, de nombreuses initiatives (portées aussi bien par des associations que par des institutions publiques ou entreprises privées) proposent d'intégrer des critères de frugalité [8] dans la conception et le déploiement des systèmes d'IA, afin de réduire l'impact environnemental de ces outils et donc aussi indirectement leur coût.

Le Nouvel Algorithme de Reconnaissance des VALeurs d'indicateurs (NARVAL) présenté dans ce rapport cherche à s'inscrire dans cette démarche. Son objectif est d'extraire des valeurs d'indicateurs codifiés sur les performances des services publics d'eau et d'assainissement en France, à partir de rapports réglementaires au format PDF fournis chaque année par les collectivités ou leurs délégataires. Jusqu'à présent, la saisie de ces 39 indicateurs est faite par les collectivités sur le site de l'observatoire national des services publics d'eau et d'assainissement (SISPEA),[11] soit manuellement, soit par l'envoi d'un fichier d'import en masse de données, et ce au plus tard le 15 octobre de l'année N pour les données de l'année $N-1$. Toutefois, cette tâche de saisie fastidieuse n'est pas toujours effectuée et dans les faits, les valeurs de ces indicateurs SISPEA s'avèrent manquantes dans la base pour près de la moitié des services (soit environ 20% de la population française non couverte). NARVAL cherche donc à faciliter le travail de saisie des collectivités, pour *in fine* enrichir la base. D'un point de vue technique, le défi est de parvenir à concevoir un outil à la fois suffisamment efficace (au regard des besoins SISPEA), économe en ressources de calcul (au regard des impacts environnementaux et des coûts financiers), et respectueux des principes de souveraineté numérique. Pour ce faire, NARVAL combine des méthodes heuristiques légères et des petits modèles de langage (SLM) ouverts. Notre approche repose également sur une forte intégration des connaissances métier (SISPEA) spécifiques au problème traité : nature et statistique des indicateurs recherchés, structures typiques d'éventuels tableaux récapitulatifs dans les rapports, différents libellés pour chaque indicateur, ...

Dans la suite de ce rapport, nous commençons par analyser brièvement en section II les documents PDF sources afin de cibler l'objectif de NARVAL. Plusieurs hypothèses sont formulées pour simplifier le problème à traiter et orienter le choix de l'approche. Nous décrivons ensuite en section III la pipeline de NARVAL, étape par étape. Dans la section IV, nous présentons les corpus de PDFs utilisés pour l'évaluation puis en section V, nous discutons des résultats obtenus. L'impact carbone de NARVAL est également évalué. Enfin, plusieurs pistes d'amélioration sont discutées en section VI. Nous concluons en section VII.

II. DÉFINITION DE L'OBJECTIF

A. Analyse préliminaire des rapports PDFs

Une première analyse des rapports sur le prix et la qualité des services (RPQS) fournis par les collectivités montre que leur forme est très variée. Tout d'abord, ils peuvent contenir à la fois du texte, des tableaux et des images. Parfois, tout le rapport PDF provient d'un document numérisé (chaque page est une image). De plus, les RPQS n'adoptent pas tous la même mise en page : certains sont proches du modèle de RPQS proposé par SISPEA[12] incluant un tableau récapitulatif, d'autres s'en inspirent en le modifiant ici et là, beaucoup suivent un format libre incluant ou non des tableaux, mentionnant ou non les codes des indicateurs. La longueur des RPQS est aussi très variable, de quelques pages à plus d'une centaine. De façon générale, même quand un tableau récapitulatif existe dans le PDF, il ne contient pas nécessairement toutes les valeurs des indicateurs. Certaines peuvent être indiquées ailleurs dans le texte ou dans des petits tableaux, d'autres ne sont tout simplement pas mentionnées dans le PDF. Enfin, les RPQS des « petites » et « grandes » collectivités sont à distinguer. Pour le comprendre, précisons que dans le référentiel SISPEA, une collectivité regroupe un ou plusieurs services (aussi appelés entités de gestion) d'eau potable ou d'assainissement, chaque service gérant un périmètre technique d'une ou plusieurs communes. Dans la suite, on parlera de « petites » collectivités pour désigner toutes celles constituées d'un seul service (pour une compétence donnée, eau ou assainissement) et d'une seule commune. Toutes les autres seront qualifiées de « grandes » collectivités. Dans les RPQS des « grandes » collectivités, les indicateurs ne sont pas toujours donnés à l'échelle attendue des services mais parfois à celles des communes. A l'inverse, dans les RPQS des « petites » collectivités, l'échelle de la collectivité se confond avec celle du service et de la commune et donc pour chaque indicateur recherché, une seule valeur (au plus) est renseignée dans le rapport PDF.

Aux RPQS fournis par les collectivités s'ajoutent les rapports annuels des délégataires (RAD) rédigés par des délégataires privés spécialisés dans la gestion de l'eau (Agur, Sodego, Aqualter, Saur, Suez, Veolia, etc). Les RADs se différencient des RPQS sur plusieurs points. D'une part, ils semblent toujours contenir un tableau récapitulatif des indicateurs réglementaires SISPEA. D'autre part, chaque délégataire a construit son propre *template* de RAD, commun à plusieurs collectivités et années d'exercice, de sorte que les formats des RADs sont moins variés que ceux des RPQS. Enfin, certains indicateurs ne sont jamais renseignés dans les RADs alors qu'ils peuvent l'être dans les RPQS.

B. Hypothèses simplificatrices

L'analyse précédente montre qu'il est complexe de vouloir s'attaquer de front au problème global, à savoir l'extraction de tous les indicateurs SISPEA d'un RPQS ou RAD quelconque. On fait donc le choix (1) de ne pas traiter les rares RPQS/RAD scannés afin de ne pas avoir à faire de l'océrisation, (2) de ne pas traiter les images dans les PDFs, (3) de ne pas traiter les RPQS/RAD correspondant à plusieurs compétences, (4) de se concentrer dans un premier temps sur les RPQS/RAD d'assainissement collectif (on aurait pu tout aussi bien commencer par ceux traitant de l'eau potable) et (5) de considérer uniquement les RPQS/RAD de « petites » collectivités avec un seul service (d'assainissement collectif) et une seule commune.

L'hypothèse (5) simplifie grandement l'objectif car dans les RPQS ou RAD de « petites » collectivités, les indicateurs sont directement donnés à l'échelle attendue. Il s'avère aussi que ce sont surtout les « petites » collectivités qui ne saisissent pas leurs indicateurs sur SISPEA, leurs moyens et leurs effectifs étant limités. Enfin, puisque le transfert des compétences du niveau communal vers le niveau intercommunal n'est plus obligatoire à ce jour, il reste tout-à-fait pertinent de développer une solution technique facilitant la saisie des données sur SISPEA pour ces « petites » collectivités.

C. Contraintes techniques

Le développement de NARVAL a été pensé en se fixant deux contraintes majeures, déjà mentionnées en introduction. La première concerne l'usage exclusif d'outils dits ouverts. En plus d'être gratuits, ils évitent le transfert de données vers des serveurs tiers et donc les problèmes inhérents en termes de transparence et de gouvernance. La seconde contrainte est le respect des principes de l'IA frugale, dans un souci de sobriété numérique mais aussi tout simplement par pragmatisme, nos moyens de calcul étant limités.[13] Ces principes nous invitent à penser autrement, à faire « mieux » avec « moins » : la recherche de l'efficacité est faite en travaillant sur l'optimisation, la parcimonie et l'intégration fine des connaissances métier, en évitant de recourir à des méthodes génériques et performantes mais lourdes. Ces deux contraintes ont guidé la conception technique de NARVAL tout au long de son développement.

III. ARCHITECTURE DE NARVAL

Les principales briques de NARVAL sont schématisées sur la figure 1 et décrites ci-après. NARVAL prend en entrée un rapport RPQS/RAD au format PDF et retourne la liste des valeurs des indicateurs au format CSV. Comme expliqué dans la section II B, seuls certains types de PDFs sont traités à ce jour. NARVAL essaie d'abord d'extraire les valeurs d'indicateurs à partir de certains tableaux dits récapitulatifs – s'ils existent – à l'aide de l'outil noté `Table Extractor`. Les indicateurs qui n'ont pas pu être extraits sont ensuite recherchés dans le texte et les éventuels autres tableaux, en interrogeant un modèle de langage noté génériquement `SLM` (pour Small Language Model). L'architecture présentée ci-dessous a été optimisée au fil des mois, en incrémentant de multiples versions mais en conservant toujours cette même approche. Nous présentons ici la version `benchmark_32` donnant en moyenne les meilleurs résultats (voir section V). La version `benchmark_table_32`, aussi discutée dans ce rapport, est la même sans l'étape d'appel au `SLM`. En effet, pour évaluer l'apport du `SLM`, nous avons comparé la pipeline dite *complète*, incluant toutes les étapes, à une pipeline dite *simplifiée*, se limitant au `Table Extractor`.

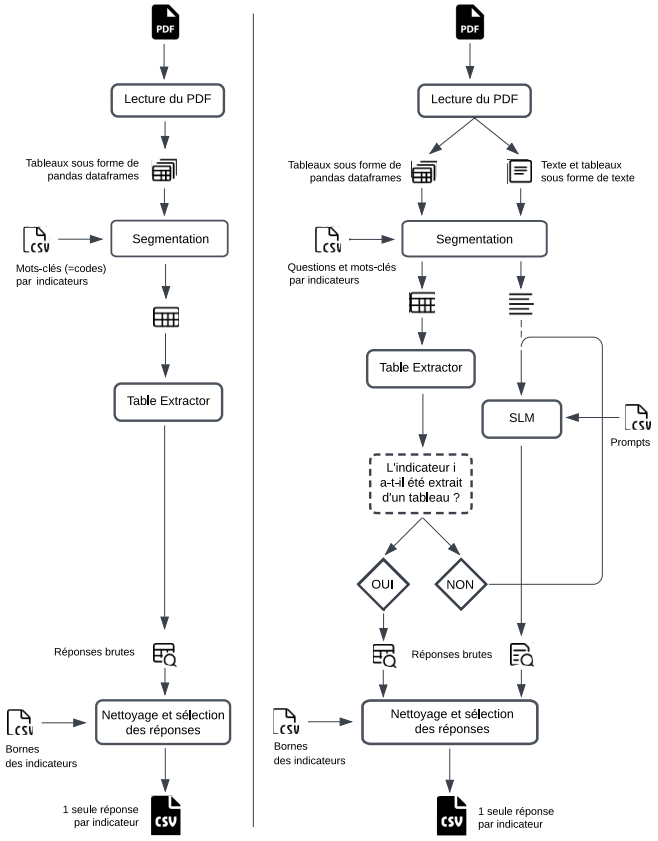

FIGURE 1. Schéma illustrant les principales étapes de NARVAL. A gauche : pipeline simplifiée (celle par exemple du `benchmark_table_32`) ; les indicateurs sont extraits uniquement des tableaux dits récapitulatifs par le `Table Extractor`. A droite : pipeline complète (celle par exemple du `benchmark_32`) ; les valeurs des indicateurs non-extraits par le `Table Extractor` sont cherchées par le `SLM`.

A. Etape 1 : lecture du rapport PDF

La première étape de NARVAL consiste à extraire le texte et les tableaux d'un PDF, les images n'étant pas traitées à ce jour.

Extraire le texte et les tableaux au format texte – On utilise la librairie python `PyPDF2`[14] pour extraire le texte et les tableaux page par page, comme du texte (à donner au `SLM`). Les tableaux ne sont pas mis en forme ici. L'extraction est intelligible mais n'est pas parfaite[15] : il y a beaucoup de coquilles, mots coupés, lettres manquantes etc ... Toutefois, les expériences menées avec NARVAL montrent que le `SLM` est capable de comprendre ce type de texte imparfait.

Extraire les tableaux en dataframes – On utilise également le package Python `PDFPlumber`[16] pour extraire les tableaux sous forme de `Pandas` DataFrame.

Détecter la table des matières – Pour chaque PDF, on extrait les numéros des pages contenant la table des matières, en combinant une recherche par mots-clés et une analyse de la structure des pages. Bien que simpliste, cette approche fonctionne pour 44 des 45 PDFs testés.[17]

Détecter si un PDF est un RAD – Enfin, par une simple recherche de mots-clés dans la première page d'un PDF, on détecte si un rapport PDF est un RAD ou non. Cette approche trop basique devrait être améliorée. Tout rapport PDF détecté comme RAD est bien un RAD mais certains RADs ne sont pas identifiés comme tels.

B. Etape 2 : segmentation du rapport PDF

Le contenu d'un rapport PDF ne peut pas être donné d'un bloc au `SLM` car sa taille de contexte est limitée et la mémoire GPU aussi. Il est donc nécessaire de segmenter le PDF : pour chaque question (à poser au `SLM` pour obtenir une réponse quant à la valeur d'un indicateur), on identifie les pages pertinentes dans le PDF pour y répondre ; de même pour chaque indicateur, on identifie les tableaux à soumettre au `Table Extractor`. Pour ce faire, une simple approche par mots-clés est utilisée.

Liste de questions et mots-clés – Il est nécessaire au préalable d'identifier pour chaque indicateur une liste de questions pertinentes, qui seront intégrées aux prompts soumis au `SLM` (voir section III D). De même, il est nécessaire d'associer à chacune de ces questions un mot-clé qui sera utilisé pour repérer la ou les page(s) associées dans un PDF. Cette liste de questions et mots-clés doit être établie à partir de la définition des indicateurs SISPEA (voir annexe A) et de la manière dont ils sont nommés en général dans les RPQS/RAD.

Code	Question	Mot-clé
D201.0	Quel le nombre d'habitants desservis par le réseau d'assainissement collectif (D201.0)	habitant
D201.0	Quelle est la valeur de l'indicateur D201.0	D201.0
P252.2	Quel est le nombre de points de curage fréquents du réseau (P252.2)	curage
P252.2	Quel est le nombre de points noirs du réseau (P252.2)	noir
P252.2	Quelle est la valeur de l'indicateur P252.2	P252.2

TABLE I. Exemples de questions posées au `SLM` pour deux indicateurs, avec les mots-clés associés utilisés à l'étape de segmentation. Chaque question est complétée par la chaine de caractères « en 2021 ? » ou « en 2022 ? » etc ... selon l'année d'exercice du rapport.

En comparant les erreurs de NARVAL pour différentes versions de cette liste, on parvient à l'améliorer progressivement. Cette étape purement empirique est longue et fastidieuse mais elle est indispensable pour garantir de bonnes performances. En suivant cette approche, nous avons par exemple constaté qu'il est préférable pour les indicateurs de conformité (P203.3, P204.3, P205.3, P254.3) de poser une seule question non ambiguë (du type « Quelle est la valeur de l'indicateur P254.3 en 2021 ? ») car si d'autres questions plus spécifiques sont posées, le `SLM` a tendance à confondre ces quatre indicateurs. En procédant ainsi, on limite le nombre de réponses chiffrées incorrectes (moins de faux positifs) mais on ferme la possibilité d'extraire la valeur d'un indicateur dont le code n'apparaît pas explicitement dans le PDF (plus de faux négatifs). Un extrait de la liste optimisée est présenté dans la table I.

Identifier les pages pertinentes pour le `SLM` – Pour chaque question, on identifie le numéro des pages du PDF contenant le mot-clé associé (au singulier ou pluriel). La table des matières (identifiée à l'étape 1) est exclue : elle contient la plupart des mots-clés mais pas la valeur des indicateurs, de sorte que si l'on interroge le `SLM` dessus, il renvoie le plus souvent le numéro de page de la section concernée. L'approche est triviale mais aussi très rapide d'exécution et il s'avère *in fine* que peu d'erreurs sont liées à une mauvaise recherche par mot-clé (si ceux-ci sont bien choisis).

Identifier les tableaux pertinents pour le `Table Extractor` – Une pré-sélection des tableaux se fait de la même manière en utilisant pour chaque indicateur uniquement son code comme mot-clé. Ainsi, pour un indicateur donné, seuls les tableaux contenant le code de l'indicateur sont soumis au `Table Extractor` à l'étape suivante.

C. Etape 3 : extraction des indicateurs des tableaux récapitulatifs

Cette étape est exécutée à l'aide de l'outil `Table Extractor` de NARVAL basé uniquement sur la librairie `Pandas`[18].

Détecter les tableaux dits récapitulatifs – Une (deuxième) sélection des tableaux est d'abord effectuée (la première étant celle de l'étape de segmentation). Seuls sont conservés les tableaux dits récapitulatifs ayant au moins une colonne contenant les codes d'indicateurs et une colonne contenant dans son titre l'année N de l'exercice du PDF (mais pas l'année $N-1$). Ainsi sont laissés de côté à ce stade les tableaux qui contiendraient par exemple les codes d'indicateurs dans une colonne et les valeurs dans une autre, sans préciser l'année, ainsi que tous les tableaux qui peuvent contenir des valeurs d'indicateurs sans leur code. Toutefois, ceux-ci sont envoyés *a posteriori* au `SLM`, sous forme de texte. A noter également que d'après cette définition, un tableau récapitulatif peut ne contenir qu'une seule valeur d'indicateur, l'adjectif « récapitulatif » faisant référence ici uniquement à la structure du tableau décrite ci-dessus.

Extraire les indicateurs – Après la mise en forme des tableaux récapitulatifs, le `Table Extractor` extrait avec `Pandas`[18] la valeur d'un indicateur en croisant simplement ligne et colonne. À ce stade, le contenu d'une cellule est extrait sous forme de chaîne de caractères, sans nettoyage. Dans certains cas, ce contenu est non chiffré (par exemple « Non disponible »), voire vide. Si l'extraction en elle-même est parfaite (`Pandas` ne se trompe pas), la détection des tableaux récapitulatifs et leur mise en forme ne le sont pas forcément : en pratique, sur les 45 PDFs testés (voir section IV), ces étapes sont aussi exécutées sans aucune erreur par le `benchmark_table_32` donnant les meilleurs résultats dans le cas de la pipeline simplifiée mais d'autres versions antérieures de NARVAL, optimisées à partir de seulement 15 de ces PDFs, gèrent mal ces étapes pour certains des 30 autres PDFs. Ainsi, à l'avenir, des erreurs pourraient subvenir pour d'autres types de tableaux pas encore rencontrés à ce jour.

D. Etape 4 : interrogation du modèle de langage

Cette étape est optionnelle. Elle n'est jamais exécutée dans le cas de la pipeline simplifiée (représentée à gauche de la Fig.1) qui extrait les indicateurs des tableaux récapitulatifs uniquement. Elle n'est pas non plus exécutée dans le cas de la pipeline complète (représentée à droite de la Fig.1) lorsqu'un indicateur a été préalablement extrait d'un tableau avec le `Table Extractor`, et ce même si la valeur extraite est non chiffrée.[19] Enfin, même si cela n'apparaît pas sur la Fig.1 (afin de ne pas l'alourdir pour un détail), cette étape 4 est également omise si un PDF est détecté comme un RAD à l'étape 1 : comme les RADs semblent toujours avoir un tableau récapitulatif fourni et sont par ailleurs très longs (typiquement une centaine de

pages), on se limite au `Table Extractor`, sans jamais interroger le `SLM`.

Dans tous les autres cas, on interroge le `SLM` en posant chaque question q sur chacune des N_q pages identifiées (pouvant contenir du texte et des tableaux sous forme de texte). Ainsi pour chaque question, on obtient une liste de N_q réponses.

Choix du modèle de langage – Deux modèles de type transformeur ont été testés pour le `SLM` : `Flan-T5-xl` de Google[20] et `Llama3-8b-Instruct` de Meta[21]. Leurs poids sont en accès libre sur la plateforme Hugging Face[22] et leur taille – respectivement 3 et 8 milliards de paramètres – les classe parmi les modèles de langage frugaux. Dès les premiers tests, le modèle `Llama3-8b-Instruct` a montré les meilleures performances,[23] notamment parce qu'il parvient à interpréter correctement les tableaux extraits comme du texte. A l'inverse, `Flan-T5-xl` a du mal à distinguer les colonnes et a tendance à renvoyer comme réponse la ligne entière correspondant à un indicateur. De plus, la taille de contexte de `Llama3-8b-Instruct` est bien plus grande (8000 tokens vs 512) ce qui permet notamment de lui donner des instructions plus précises via des prompts plus longs (voir ci-après). Ainsi, nous avons rapidement fait le choix du modèle `Llama3-8b-Instruct` et optimisé pas-à-pas la pipeline de NARVAL en fonction. Toutefois, il n'est pas exclu que des performances comparables aient pu être obtenues avec `Flan-T5-xl` (ou `Flan-T5-xxl`) en adaptant les autres étapes de la pipeline, notamment en travaillant la mise en page des tableaux, en affinant la segmentation et en modifiant les étapes de nettoyage ultérieures.

Ingénierie des prompts – Comme expliqué précédemment, les prompts pour le modèle `Flan-T5-xl` doivent être très courts car la taille du contexte est limitée pour ce modèle à 512 tokens : on ne peut donc donner comme prompt que le contenu d'une page PDF (typiquement 500 tokens), une courte question et des instructions minimales. En revanche, ce problème disparaît pour `Llama3-8b-Instruct` (bien que l'on soit tout de même toujours limité par la mémoire GPU). On peut donc fournir au modèle des instructions précises dans le prompt, indicateur par indicateur. Plusieurs versions de prompts ont été testées, certaines marchent mieux que d'autres et il n'est pas facile de l'intuiter. Un exemple de prompt est donné sur la figure 2. Il semble qu'il soit important que le *user prompt* contienne uniquement la question et le *system prompt* le contexte et les instructions. De plus, il est explicitement demandé au `SLM` de fournir une réponse concise. En effet, nous détournons ici un modèle génératif pour effectuer une tâche d'extraction, en le contraignant à générer au maximum 10 tokens par réponse. Ceci doit être précisé dans le prompt.

Absence de réponse – Dans certains cas, la valeur d'un indicateur donné n'est pas inscrite dans le PDF. Il faut donc que le `SLM` soit capable de dire qu'il ne trouve pas la réponse et éviter ainsi les hallucinations. Ce problème est traité via le prompt, en demandant au modèle de répondre `« je ne trouve pas »` s'il ne trouve pas la valeur de l'indicateur recherché.

System prompt

Tu es un assistant administratif qui répond à des questions sur les services d'assainissement collectif en France.
Tu dois extraire la valeur d'un indicateur à partir d'extraits d'un rapport sur l'assainissement collectif en {year} dans la collectivité {city}.
– – – – – – –
Instructions :
- La valeur de l'indicateur à trouver est un nombre exprimé en %.
- Ne confonds pas avec d'autres valeurs en % dans l'extrait. Si tu ne trouves pas l'indicateur recherché, réponds « je ne trouve pas ».
- Si tu ne trouves pas la réponse pour l'année {year} dans l'extrait, réponds « je ne trouve pas ».
- Si tu n'as pas assez d'information dans l'extrait pour répondre, réponds « je ne trouve pas ».
- Sois le plus concis possible. Ta réponse doit être uniquement un nombre (dans les bonnes unités) ou « je ne trouve pas ».
– – – – – – –
Extraits : {context}

User prompt

Question : Quel est le taux de renouvellement du réseau d'assainissement collectif (P253.2) en {year} ?

FIGURE 2. Exemple de prompt fourni au `SLM`, ici pour l'indicateur P253.2 et le modèle `Llama3-8b-Instruct`. Les variables {city}, {year} et {context} sont remplacées respectivement par le nom de la collectivité associée au PDF, l'année d'exercice du RPQS/RAD et le contenu d'une des pages identifiées à l'étape de segmentation.

E. Etape 5 : nettoyage et filtrage des réponses

A ce stade de la pipeline, NARVAL a extrait pour chaque indicateur i une liste de N_i valeurs possibles. Si l'indicateur a été extrait d'un (ou de X_i tableaux) via le `Table Extractor` : $N_i = X_i$ (le plus souvent $X_i = 1$). Si le `SLM` a été interrogé via M_i questions q chacune pointant vers N_{iq} pages dans le PDF : $N_i = \sum_{q=1}^{M_i} N_{iq}$. Il reste maintenant à nettoyer et filtrer ces réponses afin de n'en conserver qu'une seule : soit un nombre, soit `« je ne trouve pas »`. Ceci se fait en plusieurs étapes qui exploitent fortement notre connaissance préalable des données SISPEA et des RPQS/RADs.

(1) Nettoyage par regex – On commence par nettoyer les réponses brutes à l'aide d'expressions régulières : suppression des codes d'indicateurs (D201.0, ...), extraction de nombres à partir de texte (par exemple la chaîne de caractères « 10 580 » devient le nombre 10580), extraction des points (par exemple la réponse du `SLM` « 70 points sur un total de 120 » devient le nombre 70), suppression du texte sauf `« je ne trouve pas »`, etc. Les différents cas traités sont complétés au fur et à mesure de l'analyse des

erreurs commises par NARVAL sur un jeu de 45 PDFs (voir section V D).

(2) Suppression de valeurs pré-définies – On pré-définit un certain nombre (fini) de valeurs qu'un indicateur ne peut pas prendre, par exemple l'année courante (2021, ...) ou pour l'indicateur D204.0 (prix par m^3 par une consommation annuelle de 120m^3) le nombre 120. Toute réponse (nettoyée) du modèle qui est dans cette liste est ignorée.

(3) Suppression des hallucinations – Dans certains cas, le `SLM` invente une valeur d'indicateur qui n'est pas inscrite dans le texte. Pour limiter ce problème, on vérifie que chaque réponse (nettoyée) renvoyée par le `SLM` est bien contenue dans le texte du PDF et si non, on l'ignore. Ceci nécessite un travail par regex un peu subtil qui n'est pas parfait à ce stade et qui se généralise potentiellement mal à de nouveaux PDFs.

(4) Exclusion des valeurs aberrantes (hors bornes SISPEA) – Dans la base SISPEA, à chaque indicateur i sont associées des bornes critiques min_i^c et max_i^c au-delà desquelles une valeur d'indicateur x_i est considérée comme aberrante. Par exemple, un taux de conformité doit être nécessairement compris entre 0% et 100%. Sont également définis dans SISPEA des seuils d'alerte min_i^a et max_i^a établis à partir de l'historique des valeurs de l'indicateur i à l'échelle de la France : le plus souvent, ces seuils correspondent aux 1er et 9ème déciles de la distribution d'un indicateur. Dans NARVAL, les bornes critiques sont d'abord utilisées pour filtrer les réponses : si $x_i < min_i^c$ ou $x_i > max_i^c$, x_i est ignorée.

(5) Sélection finale – Pour sélectionner l'unique réponse finale par indicateur à partir de la liste de réponses nettoyées, on sélectionne la plus fréquente dans cette liste. Si plusieurs réponses apparaissent avec la même fréquence, on supprime celles qui sont en dehors des seuils d'alerte de SISPEA, sauf si elles sont toutes en dehors des bornes et dans ce cas on les garde toutes. S'il reste encore plusieurs réponses, on en sélectionne une aléatoirement, à défaut d'autres critères métiers de sélection. A l'issue de cette étape, à chaque indicateur est associée une unique réponse de NARVAL : soit une valeur numérique, soit `« je ne trouve pas »`.

Cette étape 5 a été mise au point de façon incrémentale, en analysant les erreurs de NARVAL, en se familiarisant peu à peu avec les données SISPEA et en améliorant le code localement ici et là. Son implémentation la rend difficilement généralisable à d'autres jeux de données, même si la méthodologie employée pourrait être transposée.

IV. JEUX TESTS DE RAPPORTS PDFS

L'évaluation de NARVAL est faite sur un jeu $\mathcal{D}_1$ de 15 RPQS, un autre jeu $\mathcal{D}_2$ de 30 RPQS ou RAD et sur la concaténation $\mathcal{D}_{1+2}$ des deux jeux, contenant 45 PDFs en tout. Ceux-ci sont téléchargeables à partir du site SISPEA.[24] Les principales propriétés de ces jeux de données sont résumées dans la table II.

	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_{1+2}$
Nombre de PDFs	15	30	45
dont RADs	0	6	6
Nombre de PDFs avec tableau(x) récapitulatif(s)	11	15	26
dont extrait(s) par le `Table Extractor` de NARVAL	9	13	22
% de valeurs d'indicateurs non indiquées dans les PDFs	$\approx 38\,\%$	$\approx 55\,\%$	$\approx 49\,\%$

TABLE II. Quelques caractéristiques des jeux de PDFs tests.

A. Choix des rapport PDFs

Pour choisir les PDFs de nos jeux tests, nous avons suivi la méthodologie suivante. Nous nous sommes limités à la compétence d'assainissement collectif et avons d'abord filtré les « petites » collectivités, c'est-à-dire celles constituées d'un seul service d'assainissement collectif et d'une seule commune. Seuls les rapports des années 2018 à 2023 ont été considérés. Nous avons ensuite exclu les RPQS scannés,[25] les télé-RPQS [26] (c'est-à-dire ceux générés automatiquement par SISPEA après saisie des indicateurs) et les PDFs traitant dans un même rapport de plusieurs compétences.[27] Dans $\mathcal{D}_1$, nous avons sélectionné des RPQS ayant un fort pourcentage ($\gtrsim 60\%$) d'indicateurs déjà saisis sur SISPEA. A l'inverse, 24 des 30 PDFs de $\mathcal{D}_2$ correspondent à des collectivités n'ayant pas fait la saisie (du moins à la date du projet). Enfin, nous n'avons pas inclus plusieurs PDFs pour la même collectivité mais diverses années d'exercice (car les PDFs sont souvent très ressemblants d'une année à l'autre pour une même collectivité). Hormis les critères ci-dessus, la sélection est aléatoire.

B. Labellisation des rapports PDFs

Pour évaluer les performances de NARVAL sur les 19 indicateurs d'assainissement collectif, il est nécessaire d'extraire pour chacun des 45 PDFs de $\mathcal{D}_{1+2}$ les valeurs des 19 indicateurs : soit une valeur numérique, soit `Null` si l'indicateur recherché n'est pas renseigné dans le PDF. Cette étape de labellisation est longue et fastidieuse mais elle doit être réalisée avec soin car elle conditionne la justesse des métriques calculées en section V et donc tout le processus d'amélioration incrémentale des versions de NARVAL basé sur ces métriques. De plus, l'idée d'utiliser les données déjà saisies dans SISPEA comme valeurs de référence ne convient pas, d'une part parce que ces données ne correspondent pas toujours aux valeurs effectivement écrites (ou effectivement manquantes) dans les rapports PDF[28] et d'autre part parce que sélectionner les RPQS de collectivités ayant fait cette saisie constitue un

biais.[29] Nous avons donc labellisé manuellement les PDFs de $\mathcal{D}_{1+2}$, en se conformant aux règles listées en annexe B. Malgré tout, dans certains cas ($\lesssim 10\%$), la valeur à extraire n'est pas claire et il est fort probable que d'autres annotateurs fourniraient des labellisations différentes. De plus, l'erreur humaine reste possible. Au terme de cette labellisation, il apparaît que seul un indicateur sur deux est effectivement renseigné dans les RPQS ou RAD de $\mathcal{D}_{1+2}$ en moyenne (voir table II). Pour deux d'entre eux, aucun indicateur n'est mentionné.

V. RÉSULTATS

A. Définition des métriques

L'efficacité de l'extraction automatique est évaluée en comparant les réponses de NARVAL aux valeurs de référence extraites manuellement des PDFs. Les réponses sont d'abord classifiées en cinq catégories :

- Vrai positif (VP) : la réponse de NARVAL est un nombre identique à celui indiqué dans le PDF,
- Vrai négatif (VN) : la réponse est `« je ne trouve pas »` et la valeur est effectivement absente du PDF,
- Faux positif de type 1 (FP1) : NARVAL renvoie une valeur numérique alors que l'indicateur n'est pas mentionné dans le PDF,
- Faux positif de type 2 (FP2) : NARVAL extrait une valeur numérique qui n'est pas celle indiquée dans le PDF,
- Faux négatif (FN) : NARVAL répond `« je ne trouve pas »` alors que la valeur de l'indicateur est indiquée dans le PDF.

On définit ensuite les métriques d'accuracy ($A_\mathcal{E}$), de précision ($P_\mathcal{E}$) et de rappel ($R_\mathcal{E}$) selon :

$$A_\mathcal{E} = \frac{N^\mathcal{E}_{VP} + N^\mathcal{E}_{VN}}{N^\mathcal{E}_{VP} + N^\mathcal{E}_{VN} + N^\mathcal{E}_{FP1} + N^\mathcal{E}_{FP2} + N^\mathcal{E}_{FN}} \quad (1)$$

$$P_\mathcal{E} = \frac{N^\mathcal{E}_{VP}}{N^\mathcal{E}_{VP} + N^\mathcal{E}_{FP1} + N^\mathcal{E}_{FP2}} \quad (2)$$

$$R_\mathcal{E} = \frac{N^\mathcal{E}_{VP}}{N^\mathcal{E}_{VP} + N^\mathcal{E}_{FN} + N^\mathcal{E}_{FP2}} \quad (3)$$

$\mathcal{E}$ désignant l'ensemble des valeurs sur lesquelles NARVAL est évalué. Ici $N^\mathcal{E}_{VP}$, $N^\mathcal{E}_{VN}$, $N^\mathcal{E}_{FP1}$, $N^\mathcal{E}_{FP2}$ et $N^\mathcal{E}_{FN}$ correspondent respectivement aux nombres de VPs, VNs, FP1s, FP2s et FNs parmi les réponses de NARVAL dans l'ensemble $\mathcal{E}$. L'accuracy fournit le pourcentage de bonnes réponses de NARVAL, tous types de réponse confondus (valeurs chiffrées ou `« je ne trouve pas »`) tandis que la précision donne le pourcentage des réponses *chiffrées* de NARVAL qui sont correctes. Enfin le rappel correspond au pourcentage des indicateurs effectivement chiffrés dans les PDFs, qui sont correctement extraits par NARVAL.

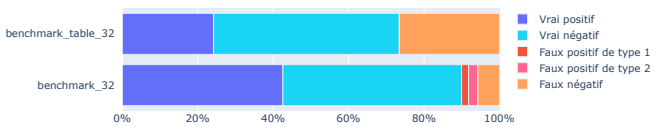

FIGURE 3. Répartition des $19 \times 45 = 855$ réponses de NARVAL pour l'ensemble des 45 PDFs du jeu $\mathcal{D}_{1+2}$.

Version	Jeu test	A_{moy}	P_{moy}	R_{moy}
`benchmark_27`	$\mathcal{D}_1$	$\approx 93.0\,\%$	$\approx 95.2\,\%$	$\approx 90.3\,\%$
`benchmark_27`	$\mathcal{D}_2$	$\approx 82.6\,\%$	$\approx 76.7\,\%$	$\approx 74.0\,\%$
`benchmark_27`	$\mathcal{D}_{1+2}$	$\approx 86.1\,\%$	$\approx 84.1\,\%$	$\approx 80.6\,\%$
`benchmark_table_28`	$\mathcal{D}_1$	$\approx 74.0\,\%$	$100\,\%$	$\approx 58.0\,\%$
`benchmark_table_28`	$\mathcal{D}_2$	$\approx 72.6\,\%$	$\approx 97.1\,\%$	$\approx 39.5\,\%$
`benchmark_table_28`	$\mathcal{D}_{1+2}$	$\approx 73.1\,\%$	$\approx 98.6\,\%$	$\approx 47.0\,\%$
`benchmark_32`	$\mathcal{D}_1$	$\approx 93.7\,\%$	$\approx 95.8\,\%$	$\approx 91.5\,\%$
`benchmark_32`	$\mathcal{D}_2$	$\approx 88.1\,\%$	$\approx 87.1\,\%$	$\approx 78.7\,\%$
`benchmark_32`	$\mathcal{D}_{1+2}$	$\approx 89.9\,\%$	$\approx 90.8\,\%$	$\approx 83.9\,\%$
`benchmark_table_32`	$\mathcal{D}_1$	$\approx 74.0\,\%$	$100\,\%$	$\approx 58.0\,\%$
`benchmark_table_32`	$\mathcal{D}_2$	$\approx 73.2\,\%$	$100\,\%$	$\approx 40.7\,\%$
`benchmark_table_32`	$\mathcal{D}_{1+2}$	$\approx 73.5\,\%$	$100\,\%$	$\approx 47.7\,\%$

TABLE III. Performance de différentes versions de NARVAL sur les jeux de test. Les performances sont évaluées par les métriques moyennes d'accuracy (A_{moy}), précision (P_{moy}) et rappel (R_{moy}). Les `benchmark_27` et `benchmark_table_28` ont été optimisés sur les 15 PDFs du jeu $\mathcal{D}_1$ tandis que les `benchmark_32` et `benchmark_table_32` l'ont été sur l'ensemble des 45 PDFs du jeu $\mathcal{D}_{1+2}$.

Ces métriques peuvent être calculées PDF par PDF, indicateur par indicateur ou en moyenne sur l'ensemble des 19 indicateurs et des 15, 30 ou 45 PDFs des jeux tests $\mathcal{D}_1$, $\mathcal{D}_2$ et $\mathcal{D}_{1+2}$.

B. Métriques moyennes

La répartition des réponses de NARVAL en vrais et faux positifs et négatifs est représentée sur la figure 3 pour la pipeline simplifiée (`benchmark_table_32`, en haut) et la pipeline complète (`benchmark_32`, en bas). Ici, les $19 \times 45 = 855$ réponses de NARVAL pour l'ensemble des 45 PDFs du jeu $\mathcal{D}_{1+2}$ sont considérées. Lorsque les indicateurs sont extraits uniquement des tableaux dits « récapitulatifs » (`benchmark_table_32`), seuls $\approx 24\%$ des réponses de NARVAL sont des vrais positifs (bleu foncé) alors que $\approx 51\%$ des indicateurs sont renseignés dans les PDFs de $\mathcal{D}_{1+2}$ (voir la dernière ligne de la table II) : ainsi, moins d'un indicateur chiffré sur deux est extrait par NARVAL. Cette pipeline simplifiée a toutefois l'avantage de ne générer aucun faux positif ; toutes les erreurs de NARVAL ($\approx 27\%$ des réponses) sont des faux négatifs (en orange). Ils s'expliquent simplement par le fait qu'ici le texte des PDFs n'est pas exploré et seul un certain

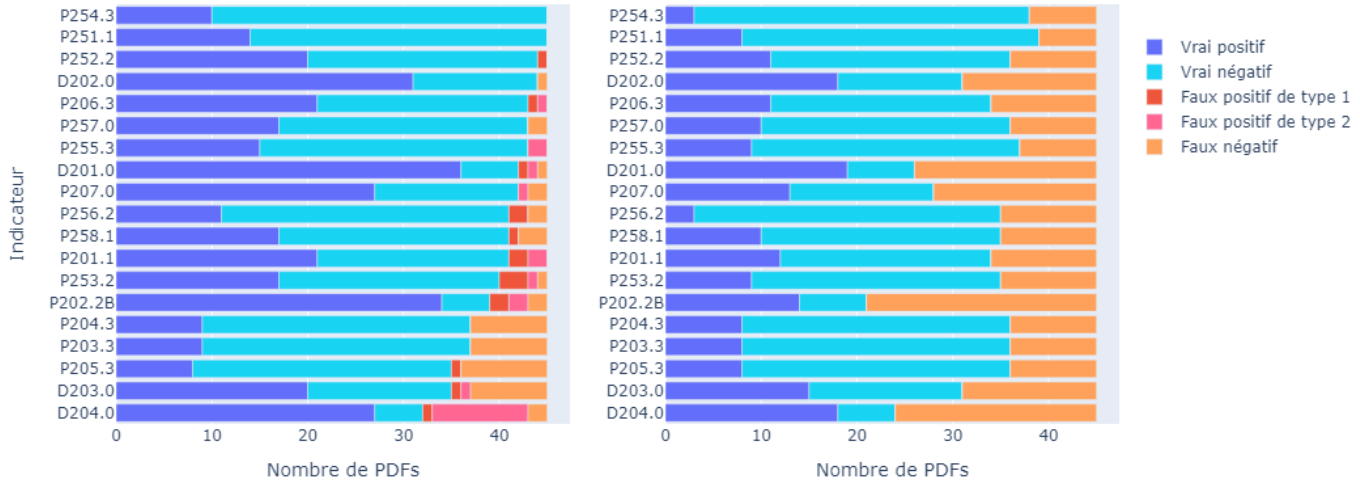

FIGURE 4. Répartition pour chaque indicateur des 45 réponses de NARVAL (correspondant aux 45 PDFs du jeu $\mathcal{D}_{1+2}$). Les résultats sont montrés pour le `benchmark_32` (à gauche) et le `benchmark_table_32` (à droite).

type de tableaux est investigué. A l'inverse, dans le cas de la pipeline complète (`benchmark_32`) faisant appel au `SLM`, près de $\approx 43\%$ des réponses de NARVAL sont des vrais positifs (bleu foncé), soit $\approx 84\%$ des $\approx 51\%$ attendus. Ceci s'accompagne toutefois d'une émergence de faux positifs (en rose et rouge) représentant $\approx 4.3\%$ des réponses.

Les métriques correspondantes d'accuracy, précision et rappel sont données dans la table III pour les `benchmark_32` et `benchmark_table_32` qui ont été optimisés sur le jeu $\mathcal{D}_{1+2}$. Sont aussi présentés les résultats pour les `benchmark_27` et `benchmark_table_28` qui, eux, ont été optimisés sur le jeu $\mathcal{D}_1$ contenant seulement 15 des 45 PDFs de $\mathcal{D}_{1+2}$. Concernant les `benchmark_32` et `benchmark_table_32`, les conclusions du paragraphe précédent issues de l'analyse de la figure 3 peuvent se reformuler ainsi : 100% des réponses chiffrées du `benchmark_table_32` sont correctes pour $\mathcal{D}_{1+2}$ (précision de 100%) mais seules $\approx 48\%$ des réponses renseignées dans les PDFs sont (correctement) extraites (rappel de $\approx 48\%$) ; pour le `benchmark_32`, le rappel grimpe à $\approx 84\%$ mais la précision chute à $\approx 91\%$.

L'analyse de la table III permet également de mettre en évidence l'épineux problème de généralisation de NARVAL à de nouveaux PDFs. On constate par exemple que le `benchmark_27` optimisé sur $\mathcal{D}_1$ affiche une précision de $\approx 95\%$ et un rappel de $\approx 90\%$ sur ce même jeu $\mathcal{D}_1$. En revanche, lorsqu'il est exécuté sur les 30 autres PDFs de $\mathcal{D}_2$, la précision et le rappel chutent respectivement à $\approx 77\%$ et $\approx 74\%$. C'est seulement en analysant les erreurs faites sur $\mathcal{D}_2$ que nous avons pu corriger NARVAL jusqu'à obtenir les résultats du `benchmark_32`. Il en est de même pour la pipeline simplifiée (`benchmark_table_28` et `benchmark_table_32`) car certains tableaux dits « récapitulatifs » de $\mathcal{D}_2$ ont un format légèrement différent de ceux de $\mathcal{D}_1$, nécessitant une mise à jour du `Table Extractor` (identification des colonnes utiles, nettoyage des tableaux). Ainsi, rien ne garantit que nos meilleures métriques obtenues avec les `benchmark_32` et `benchmark_table_32` sur $\mathcal{D}_{1+2}$ restent valables pour un autre jeu de PDFs.

C. Métrique par indicateur et par PDF

Il est aussi intéressant d'analyser les performances de NARVAL pour chacun des 19 indicateurs d'assainissement collectif et pour chacun des 45 rapports PDF de $\mathcal{D}_{1+2}$. Les résultats sont discutés ci-dessous pour le `benchmark_32` et le `benchmark_table_32`, présentant respectivement les meilleures performances moyennes par rapport aux autres versions testées dans le cas de la pipeline complète et de la pipeline simplifiée. Néanmoins, certaines de ces autres versions conduisent à de meilleures métriques d'accuracy, précision ou rappel pour certains PDF ou certains indicateurs. En effet, il arrive qu'en voulant corriger les erreurs pour l'un, on en crée de nouvelles pour l'autre.

La figure 4 montre que les performances de NARVAL varient beaucoup d'un indicateur à l'autre. Avec le `benchmark_32`, deux indicateurs (P254.3 et P251.1) sont parfaitement extraits. A l'inverse, le taux de faux positifs est particulièrement important pour l'indicateur D204.0 (prix TTC du service par m3), ce qui s'explique par la tendance du `SLM` à confondre avec d'autres prix en €/m3 mentionnés dans les PDFs, notamment celui de l'année $N-1$. On remarque également un nombre important de faux négatifs pour les indicateurs de conformité P203.3, P204.3 et P205.3, dûs à des modifications[30] volontaires au fil des versions permettant de limiter le nombre de faux positifs (dont l'impact est bien plus délétère). L'analyse de la figure 4 permet d'identifier rapidement les indicateurs pour lesquels NARVAL doit être encore amélioré. A défaut, on pourra décider de se limiter à la pipeline simplifiée pour extraire certains indicateurs : par exemple, en utilisant le `benchmark_table_32` au lieu

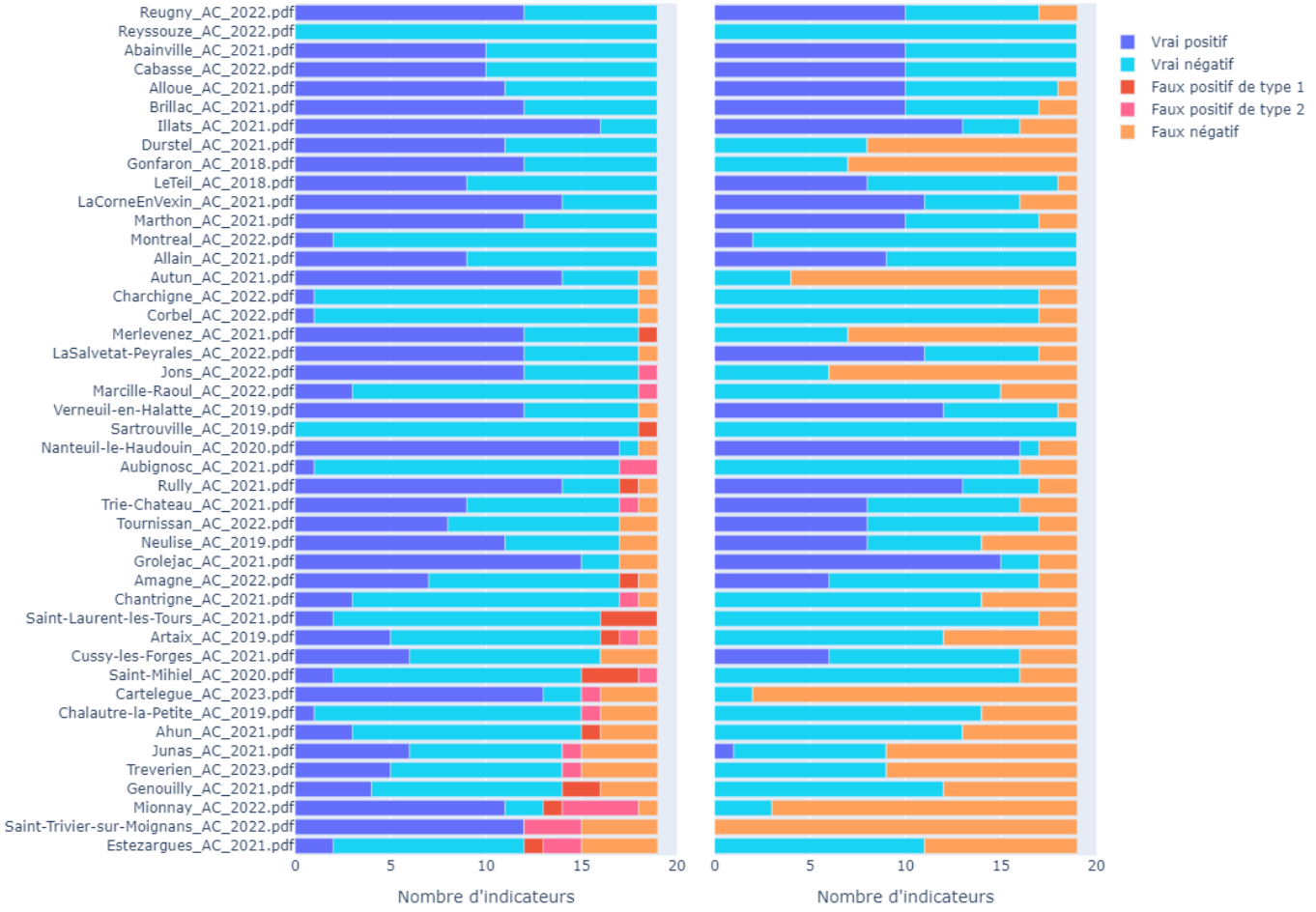

FIGURE 5. Répartition des 19 réponses de NARVAL pour chacun des 45 PDFs du jeu $\mathcal{D}_{1+2}$. Les résultats sont montrés pour le `benchmark_32` (à gauche) et le `benchmark_table_32` (à droite).

du `benchmark_32` pour D204.0, on extrait seulement 18 vrais positifs (au lieu de 27) mais on évite 11 faux positifs (sur 45 valeurs en tout).

Le même type d'analyse est reproduit PDF par PDF sur la figure 5. Les résultats montrent également une grande variabilité d'un PDF à l'autre. On remarque en particulier que l'accuracy pour le `benchmark_32` varie de $\approx$ 63% à 100% selon les PDFs, avec 100% d'accuracy pour 14 PDFs sur 45. Sans surprise, les PDFs pour lesquels NARVAL renvoie le plus de faux positifs (ceux des communes de Saint-Laurent-les-Tours, Saint-Mihiel, Mionnay, Saint-Triviers-sur-Moignans et Estezargues) n'ont aucun tableau dit « récapitulatif » lu par le `Table Extractor`. Parmi ces PDFs, deux partagent le modèle de mise en page d'un même prestataire et certaines phrases contenant des formules sont mal extraites à l'étape 1 de la pipeline (si bien que le `SLM` renvoie la réponse pour l'année $N-1$ au lieu de N). Deux autres PDFs ont peu de valeurs d'indicateurs effectivement renseignées et le `SLM` fait la confusion avec d'autres valeurs, souvent à cause d'une tournure de phrase dans le PDF ressemblant à la question posée au `SLM`.

D. Analyse des erreurs

Les résultats présentés ci-dessus montrent que la meilleure version de NARVAL correspondant au `benchmark_32` présente un taux d'erreur d'environ 10.1% sur les 45 PDFs de $\mathcal{D}_{1+2}$. Ainsi, sur $19 \times 45 = 855$ valeurs à extraire, 86 sont fausses, soit parce que la valeur numérique extraite ne correspond pas à celle indiquée ou non dans le PDF (faux positifs), soit parce que NARVAL renvoie `« je ne trouve pas »` alors que la valeur de l'indicateur est bien indiquée dans le PDF (faux négatifs). Pour comprendre l'origine de ces erreurs et pouvoir les corriger, il est nécessaire d'aller les regarder une par une, PDF par PDF. Les principales conclusions de notre analyse pour le `benchmark_32` sont listées dans la table IV. Les erreurs sont d'origine très diverse et 5 d'entre elles (parmi 86) sont associées à une difficile labellisation des PDFs : dans ces cas-là, on s'attend à ce que plusieurs humains non-experts donnent en général des réponses différentes. Ces 5 erreurs sont donc irréductibles (aux fluctuations statistiques près), à moins de clarifier les règles de labellisation (annotateurs experts) et de modifier en

Type d'erreur	Nombre	Origine
Faux négatif	49	Le mot-clé utilisé à l'étape de segmentation n'apparaît pas dans le PDF [a]
		L'indicateur est marqué « conforme » ou coché « oui » dans le PDF. [b]
		La réponse est extraite d'un tableau dit récapitulatif par le `Table Extractor` et est non chiffrée (« sans objet », « non estimable », « oui », ...) : le `SLM` n'est pas interrogé alors que la réponse est ailleurs dans le PDF, ou « oui » correspond à 100%.
		La segmentation par mot-clé renvoie la page N alors que la réponse est écrite en page $N+1$.
		NARVAL cherche dans la bonne page mais ne trouve pas, à cause d'une mauvaise extraction du texte ou d'une formulation de question non adaptée au texte.
		La réponse du `SLM` est tronquée (1 seule erreur). [c]
		Le PDF est un RAD : NARVAL n'a cherché que dans les tableaux dits récapitulatifs alors que la réponse est ailleurs dans le PDF.
		Le `SLM` renvoie une réponse incohérente qui est ensuite filtrée à l'étape de nettoyage.
		La réponse est dans un tableau au format image, non extrait à l'étape 1 de la pipeline.
		Mauvais traitement des hallucinations.
		Coquilles dans le PDF ou ambiguité quant à la valeur réelle de l'indicateur.
Faux positif de type 1	16	NARVAL renvoie la réponse pour l'année $N-1$ alors que celle pour l'année N n'est pas indiquée dans le PDF.
		NARVAL renvoie une valeur exprimée dans la même unité que l'indicateur recherché. [d]
		NARVAL semble extraire « au hasard » une réponse du PDF, sans respect de l'unité
		Mauvais traitement des hallucinations.
		Ambiguité quant à la valeur réelle de l'indicateur.
Faux positif de type 2	21	NARVAL renvoie la réponse pour l'année $N-1$ au lieu de N.
		La segmentation par mot-clé renvoie la page N alors que la réponse est écrite en page $N+1$.
		Le `SLM` renvoie une valeur qui est dans le PDF mais n'est pas la bonne. [e]
		Mauvais traitement des hallucinations.
		Problème de sélection entre réponses (choix aléatoire).
		Labellisation incertaine ou ayant nécessité un raisonnement.

[a] Ceci est fréquent pour les indicateurs de conformité P203.3, P204.3 et P205.3 car on a volontairement simplifié les questions et mot-clés pour ces indicateurs afin d'éviter des faux positifs.

[b] Ceci concerne les indicateurs de conformité. Des tests ont été effectués en ajoutant dans les prompts une instruction du type « si l'indicateur est indiqué comme conforme, réponds 100% » mais elle a été retirée afin d'éviter des faux positifs.

[c] Le `SLM` est limité à 10 tokens en sortie et il lui est demandé dans le prompt de générer une réponse concise. On pourrait augmenter le nombre de tokens générés mais cela augmenterait le temps de calcul.

[d] Parfois sa confusion est compréhensible, parfois non car à part l'unité, la valeur extraite n'a aucun rapport avec l'indicateur recherché.

[e] Parfois la bonne réponse est aussi proposée par NARVAL mais avec une moindre fréquence et donc écartée.

TABLE IV. Analyse des 86 erreurs du `benchmark_32` sur les 45 PDFs du jeu $\mathcal{D}_{1+2}$.

fonction les prompts soumis au `SLM`.

C'est en faisant ce travail d'analyse des erreurs des différentes versions de NARVAL au fil des semaines que nous avons pu améliorer la pipeline de façon incrémentale. Le `benchmark_32` reste améliorable mais la diversité des erreurs nécessite d'agir à petit pas sur les différentes étapes de la pipeline. Plusieurs pistes de solution sont proposées dans la section VI.

E. Temps de calcul et impact carbone

Pour exécuter la pipeline complète de NARVAL avec appel au `SLM`, il est nécessaire de disposer de moyens de calculs suffisants, avec typiquement un GPU $\gtrsim 16$ Go. En revanche, si le `SLM` n'est pas interrogé et que les indicateurs sont extraits uniquement des tableaux par le `Table Extractor`, NARVAL peut être exécuté sur un PC de bureau basique. Nous avons comparé les temps de calcul pour ces deux approches lorsque nous utilisons respectivement (1) un PC portable DELL avec un processeur i5-13600H, 2.80 GHz, 16 Go de RAM, sans GPU, et (2) la plateforme Datalab du SSPCloud,[31] en demandant un GPU de 32-64 Go et une centaine de CPUs. Ces temps de calcul varient beaucoup d'un PDF à l'autre, selon le nombre de pages du PDF et le nombre de tableaux récapitulatifs. Les temps moyens τ_{laptop} et τ_{SSPCloud} pour un PDF du jeu $\mathcal{D}_{1+2}$ sont donnés dans la table V.

Nous avons également estimé l'impact carbone de NARVAL avec le package CodeCarbon.[32] Celui-ci estime la consommation énergétique de la machine utilisée pour l'exécution du code et en déduit une approximation de

	τ_{laptop}	τ_{SSPCloud}	$CO_{2,eq}$
`benchmark_table_32`	≈ 8 sec.	-	≈ 5 mg
`benchmark_32`	-	≈ 2 min.	≈ 1 g

TABLE V. Temps de calcul et impact carbone moyens de NARVAL pour un PDF du jeu $\mathcal{D}_{1+2}$.

l'impact carbone, connaissant le mix énergétique du pays où la machine est située. L'estimation est de 1 g $CO_{2,eq}$ en moyenne par PDF pour une exécution du `benchmark_32` sur le SSP Cloud (situé en France). Ainsi une exécution sur 10000 PDFs par an émettrait environ 10 kg $CO_{2,eq}$ (soit 0.1% de l'empreinte carbone annuelle moyenne d'un citoyen français en 2021). Cette estimation ne prend pas en compte l'impact carbone lié à la fabrication de la machine, ainsi que les autres impacts environnementaux (consommation d'eau, de ressources abiotiques, etc). Ils pourraient être estimés avec d'autres outils, comme l'API Boavizta.[33]

VI. PERSPECTIVES

La version actuelle de NARVAL correspond à un projet exploratoire. Beaucoup d'améliorations pourraient être apportées, soit en conservant l'approche actuelle et en essayant de l'améliorer brique par brique, soit en changeant de paradigme et en explorant de nouvelles pistes. Quelle que soit l'approche choisie, il sera également nécessaire de généraliser NARVAL aux indicateurs d'eau potable et d'assurer un suivi des performances de NARVAL.

A. Optimisation de la pipeline existante

L'analyse des erreurs en section V D a montré que les erreurs commises par NARVAL sont d'origine multiple. A moins de changer complètement de paradigme (voir la section suivante), ces erreurs ne pourront être corrigées en travaillant sur un unique levier. Au contraire, toutes les briques de la pipeline doivent être optimisées. Quelques pistes sont proposées ci-dessous.

Améliorer la segmentation – Plusieurs erreurs de NARVAL sont dûes à une mauvaise segmentation des PDFs : soit le mot-clé pointe vers une page non pertinente, soit le mot-clé apparait trop souvent dans le PDF et il est alors nécessaire de sélectionner une réponse parmi un grand panel. Pour résoudre ce problème, on pourrait commencer par segmenter le PDF par bloc et non par page. Ainsi, au lieu d'extraire la page N contenant un mot-clé donné, on pourrait extraire la portion de texte qui commence une ou deux phrases avant le mot-clé et se termine environ une longueur de page après. Ceci permettrait de résoudre les erreurs dûes au fait que parfois un mot-clé pointe vers la page N alors que la réponse est en page $N + 1$. On pourrait aussi essayer d'exploiter d'autres packages python pour détecter les sections du PDF et extraire le texte section par section. Une autre approche, plus complexe, consisterait à baser la segmentation du PDF non pas sur une recherche de mots-clés mais sur une recherche sémantique, à la manière d'un modèle RAG (Retrieval Augmented Generation). L'idée serait d'identifier précisément les phrases (et non les pages) d'un PDF informant sur la valeur d'un indicateur, en combinant des calculs de *sentence embedding* et de *sentence similarity*, puis de donner ces phrases comme unique contexte au `SLM` pour un indicateur donné. Par construction, le modèle fournirait une seule réponse par indicateur, évitant ainsi l'étape imparfaite de sélection *a posteriori*.

Tester d'autres modèles de langage – Il est évident que d'autres petits modèles de langage ouverts pourraient être testés, comme par exemple `Ministral-8b-Instruct` adapté pour traiter du texte en français, `OLMo2-7B-Instruct`, et bien d'autres. Il serait intéressant aussi de tester un modèle de taille intermédiaire,[34] tel que `LLama3.1-70b-Instruct`, et d'évaluer le gain en performance par rapport au surcoût environnemental. Une autre approche consisterait à utiliser un modèle extractif – par opposition à génératif – de type camemBERT[1,35] qui, étant donné une question et un contexte, extrait une portion du contexte comme réponse, en associant un score de confiance. En analysant ce score, on pourrait déterminer s'il existe un seuil au-dessus duquel le modèle renvoie toujours (ou presque) le bon extrait. Si oui, on pourrait imaginer de diviser l'étape 4 de la pipeline en deux : le modèle léger BERT serait utilisé dans un premier temps puis, seulement si celui-ci n'est pas sûr (score de confiance en-dessous d'un seuil à déterminer), un modèle plus lourd comme `LLama3-8b-Instruct` serait appelé.

Faire du fine-tuning des modèles de langage – Le vocabulaire et les tournures de phrases utilisés dans les RPQS et RAD sont conformes aux besoins de SISPEA mais ne correspondent pas au langage courant, mathématique ou informatique utilisés dans les jeux de données d'entraînement de `LLama3-8b-Instruct`. On s'attend donc à une amélioration des performances de NARVAL après *fine-tuning* des poids de ce modèle. Toutefois, on s'attend aussi à ce que la réalisation de cette tâche de *fine-tuning* soit chronophage car il faut préparer des données d'entraînement labellisées (non utilisées pour l'évaluation de NARVAL) puis ré-entraîner le modèle sur celles-ci.

Améliorer le `Table Extractor` – A ce jour, le `Table Extractor` a une précision de 100% sur le jeu $\mathcal{D}_{1+2}$ mais seuls certains tableaux dits récapitulatifs sont traités : avec une colonne A contenant les codes des indicateurs et une colonne B contenant leurs valeurs, étiquetée explicitement par l'année d'exercice. On pourrait envisager de traiter également les tableaux dont la colonne B ne mentionne pas explicitement l'année d'exercice dans son titre (mais par exemple uniquement « valeur »).[36] De même, on pourrait envisager de traiter les tableaux spécifiques aux indices de connaissance P202.2B et

P255.3 qui apparaissent dans certains RPQS. Ceci permettrait aussi d'extraire les variables intermédiaires qui sont à saisir sur SISPEA *in fine*.

Améliorer l'extraction du texte – Comme mentionné à la section III A, le texte extrait par `PyPDF2` est plein de coquilles. En général, cela ne pose pas de problème pour le `SLM` mais certains bouts de texte de certains PDFs sont trop mal extraits pour être compris (notamment des formules mathématiques mal retranscrites). Il serait donc préférable de nettoyer le texte avant de l'envoyer au `SLM`, soit en testant des librairies python de correction orthographique, soit en remplaçant `PyPDF2` par une autre librairie d'extraction de texte.

Améliorer le traitement des hallucinations – Le traitement actuel des hallucinations souffre de trois problèmes : 1) une mauvaise extraction du texte, dans certains cas (voir ci-dessus), 2) la difficulté à interpréter les espaces entre une série de nombres et 3) une mauvaise implémentation (facilement améliorable) qui procède en cherchant la présence d'une valeur chiffrée d'un indicateur dans l'ensemble des pages sélectionnées pour cet indicateur (à l'étape de segmentation) au lieu de chercher uniquement dans la page dont provient la réponse. Le modèle `LLama3-8b-Instruct` étant sujet aux hallucinations, il est important de corriger ces problèmes.

Améliorer la détection de la table des matières – A ce jour, la table des matières est détectée par une approche basique combinant recherche de mots-clés et analyse de la structure des pages. L'approche fonctionne pour 44 des 45 PDFs de $\mathcal{D}_{1+2}$ mais peut-être pas sur d'autres PDFs. On pourrait envisager de construire un classifieur de page, à base par exemple de forêts aléatoires. Cette tâche est toutefois non prioritaire.

Améliorer la détection et le traitement des RADs – Dans la version actuelle de NARVAL, on détecte si un PDF est un RAD en cherchant la présence de mots-clés dans la première page du PDF. Cette approche simpliste ne permet pas d'identifier tous les RADs et pourrait être facilement améliorée en construisant par exemple un modèle de classification supervisée. Par ailleurs, NARVAL utilise uniquement le `Table Extractor` pour extraire les valeurs d'indicateurs de RADs alors que dans certains cas (rares), ces valeurs ne sont pas indiquées dans les tableaux récapitulatifs. Comme les RADs sont souvent longs, traiter tout le document avec le `SLM` serait couteux et imprécis. On pourrait envisager de détecter et supprimer en amont le glossaire et annexes inutiles afin de réduire la taille du PDF avant d'interroger le `SLM`. Ces améliorations ne sont toutefois pas prioritaires car plusieurs organismes rédigeant les RADs ont conclu avec l'OFB un échange automatisé des valeurs d'indicateurs pour alimenter aisément SISPEA.

Améliorer la sélection des réponses – A l'étape 5 de la pipeline, une valeur finale d'indicateur doit être choisie parmi plusieurs valeurs candidates. La connaissance de seuils critiques et de seuils d'alerte définis par indicateur à l'échelle nationale est utilisée pour faire cette sélection mais on pourrait l'affiner en analysant plus finement la statistique des indicateurs. Par exemple, des seuils pourraient être définis par collectivité en exploitant l'historique des données ou par collectivités aux propriétés similaires, suite à une analyse par *clustering*. A défaut de pouvoir faire un choix raisonné entre plusieurs réponses candidates, il serait préférable de n'en sélectionner aucune (c'est-à-dire de renvoyer `« je ne trouve pas »`) au lieu d'en sélectionner une aléatoirement, afin de limiter le nombre de faux positifs. Enfin, comme mentionné ci-dessus, on pourrait également améliorer l'étape de segmentation de façon à restreindre le nombre de réponses candidates par indicateur et donc simplifier la tâche de sélection.

B. En changeant d'approche

Exploiter un modèle propriétaire – La version actuelle de NARVAL a été développée en se contraignant à utiliser des outils à la fois ouverts et frugaux. Au vu des résultats obtenus, notamment de la précision insuffisante pour une mise en production (voir Table III), il serait intéressant de voir à quel point un outil propriétaire exploitant un LLM pour la compréhension de PDFs (Claude 3.5 Sonnet, Google Document AI, Open AI GPT4 ...) performe mieux que notre solution. Il ne s'agit pas de s'affranchir complètement de la contrainte de frugalité mais d'estimer le rapport entre les bénéfices (gain de performance, facilité d'exécution sur une simple machine de bureau via une API) et les coûts (financier, environnemental, ...) d'une telle approche. De plus, le travail fourni pour développer NARVAL pourra tout de même être valorisé : *a priori* il sera toujours intéressant d'employer l'outil `Table Extractor` dans un premier temps, de sorte que l'API ne sera appelée que sur les $\approx 50\%$ des indicateurs qui ne peuvent être extraits des tableaux dits récapitulatifs ; de plus, l'écriture des prompts détaillés par indicateur bénéficiera de notre connaissance acquise sur la manière dont sont typiquement mentionnés les indicateurs dans les RPQS. L'utilisation d'une API permettra en revanche d'allonger les prompts avec des instructions complémentaires (en veillant à rester concis tout de même pour limiter les coûts) et de simplifier considérablement la pipeline, notamment les étapes de segmentation et de nettoyage, puisque tout le PDF pourra en principe être traité d'un bloc par le LLM.[37] Il sera aussi intéressant d'estimer s'il vaut mieux interroger le LLM N fois pour les N indicateurs ou s'il est préférable d'écrire un prompt géant demandant au LLM de renvoyer tous les indicateurs sous un format structuré. Si la solution développée est performante mais s'avère trop coûteuse au sens financier et/ou écologique,[38] elle pourra être utilisée pour labelliser occasionnellement des PDFs. Une autre piste intéressante consisterait à n'interroger un LLM propriétaire qu'en dernier recours, lorsque l'approche actuelle échoue à extraire une valeur (par exemple si aucune réponse n'est retenue après nettoyage ou si plusieurs ré-

ponses semblent possibles). Une telle stratégie hybride permettrait de bénéficier ponctuellement des capacités d'un modèle plus puissant, tout en limitant les appels et donc les coûts associés. La mise en œuvre de cette architecture pourrait être facilitée par des frameworks comme LangChain,[39] qui offrent des briques pour combiner différents modèles et automatiser les prompts.

Traiter les RPQS de « grandes » collectivités – La version actuelle de NARVAL ne traite que les RPQS/RAD de « petites » collectivités constituées d'un seul service et d'une seule commune. Traiter le cas générique est complexe car l'échelle à laquelle les indicateurs sont attendus dans SISPEA (celle des services) n'est pas toujours celle utilisée dans les rapports PDF (souvent celle de la commune). Ainsi, pour généraliser par exemple le `Table Extractor`, il faudra préalablement développer un outil permettant de distinguer les tableaux des différents services et communes. De même, il faudra modifier les prompts pour que les questions soient posées par service ou commune, nécessitant sans doute l'utilisation d'un LLM. Enfin, en bout de chaîne, le cas échéant, les indicateurs devront être recalculés à l'échelle d'un service connaissant les valeurs à l'échelle d'une commune.

C. Vers une mise en production de l'outil NARVAL

Généraliser à d'autres indicateurs et variables – Seuls sont traités les 19 indicateurs d'assainissement collectif (voir annexe A) dans la version actuelle de NARVAL. Ceci est suffisant pour l'étude exploratoire présentée dans ce rapport mais dans la pratique, d'autres indicateurs d'eau potable et d'assainissement non collectif doivent également faire l'objet d'une saisie sur SISPEA. De même, les variables servant au calcul de ces indicateurs doivent être saisies en principe, certaines à une échelle plus fine que celle du service (station d'épuration, station de prélèvement, ...). Généraliser NARVAL à ces indicateurs et variables ne présente pas de difficulté particulière[40] car la même approche peut être reproduite. Il faudra d'abord labelliser de nouveaux RPQS/RADs afin d'être en mesure d'évaluer la performance de NARVAL sur ces nouveaux cas. Il faudra ensuite mettre à jour la liste des questions, mots-clés associés et les instructions des prompts. Ceci nécessitera certainement plusieurs tentatives en fonction des résultats obtenus et de l'analyse des erreurs. Cette analyse permettra également d'estimer s'il est nécessaire ou non de mettre à jour l'étape 5 de nettoyage dans la pipeline.

Ajouter et automatiser des tests unitaires – Les bonnes pratiques de gestion des tests ne sont pas respectées dans la version actuelle de NARVAL. Seuls des *notebooks* ont été utilisés au fur et à mesure du développement pour tester les différentes briques de la pipeline : test de nettoyage des réponses, de suppression des hallucinations, de détection de la table des matières, ... Ces tests doivent être ré-écrits, complétés et automatisés via la mise en place d'une pipeline d'intégration continue (CI) sur gitlab.

Mettre en place un outil de monitoring – Les performances de NARVAL peuvent se dégrader lorsque celui-ci est évalué sur un jeu de PDFs statistiquement différent du jeu utilisé pour optimiser la pipeline (voir section V). Ainsi, en phase de production, il sera important de mettre en place un outil de *monitoring* pour alerter quant à la dégradation des performances de NARVAL, au risque sinon d'extraire au fil des mois des valeurs d'indicateurs de moins en moins fiables. Pour cela, il faudra labelliser un échantillon de PDFs tous les quelques mois, contrôler les métriques sur cet échantillon et éventuellement mettre à jour la pipeline si la précision et le rappel s'abaissent en dessous d'un seuil critique à définir. On pourra s'appuyer pour cela sur des frameworks comme LangChain.[39]

VII. CONCLUSION

Nous avons développé en python un outil frugal et ouvert permettant d'extraire les indicateurs SISPEA (au format CSV) à partir des rapports RPQS ou RAD (au format PDF) fournis par les collectivités ou leurs délégataires. Cet outil NARVAL est limité à ce jour aux 19 indicateurs d'assainissement collectif et ne traite que les rapports PDF des « petites » collectivités, constituées chacune d'un seul service d'assainissement collectif et d'une seule commune. Les images contenues dans les PDFs ou les pages scannées ne sont pas non plus traitées. Nous avons évalué NARVAL sur un jeu test de 45 PDFs pour deux architectures de pipeline. La première, dite simplifiée, extrait uniquement les indicateurs de certains tableaux récapitulatifs. Elle s'exécute en quelques secondes (par PDF) sur un PC de bureau basique (sans GPU). La seconde, dite complète, fait appel au modèle de langage `Llama3-8b-Instruct` pour explorer le texte et les autres tableaux. Elle nécessite un GPU $\gtrsim$ 16 Go et s'exécute alors en quelques minutes (par PDF). La pipeline simplifiée permet d'extraire en moyenne $\approx 48\%$ des valeurs renseignées dans les 45 PDFs testés, avec une précision de 100%. Autrement dit, toutes les valeurs extraites sont correctes mais moins d'un indicateur sur deux est trouvé (certains PDFs n'ayant pas de tableau récapitulatif au format attendu). A l'inverse, la pipeline complète permet d'extraire correctement $\approx 84\%$ des indicateurs en moyenne mais parmi les réponses chiffrées renvoyées par NARVAL, seulement $\approx 91\%$ sont correctes. Les résultats montrent par ailleurs une grande variabilité d'un PDF à l'autre et d'un indicateur à l'autre.

La comparaison des résultats de NARVAL pour les pipelines complète et simplifiée permet ainsi de quantifier l'apport de l'IA dans un projet concret d'analyse de texte et de tableaux issus de rapports PDFs. Ici l'appel au SLM dans la pipeline complète permet presque de doubler le nombre de vrais positifs (c'est-à-dire d'indicateurs correctement extraits) mais près de 5% des réponses de NARVAL sont alors des faux positifs (*versus* 0% pour la pipeline simplifiée). Dans le cadre du projet

NARVAL, l'émergence de ces faux positifs est particulièrement problématique car il est crucial de pouvoir garantir la fiabilité des indicateurs SISPEA extraits à partir des RPQS ou RAD. Plusieurs choix méthodologiques ont pourtant déjà été faits, tout au long du développement de NARVAL, pour minimiser le nombre de faux positifs, au détriment du rappel. Bien d'autres sont encore envisageables. Le bon compromis à trouver entre précision et rappel dépend du problème métier à traiter et il se peut que dans d'autres contextes, un outil d'extraction automatisée offrant 90% de précision apporte déjà une grande valeur ajoutée aux utilisateurs. De façon générale, notre étude illustre la pertinence d'une approche hybride, combinant expertise métier, méthodes heuristiques et intelligence artificielle, comme levier d'innovation dans l'administration publique, confrontée régulièrement à des besoins d'extraction de données structurées à partir de rapports PDFs. Une telle approche permet de se poser les bonnes questions, d'exploiter les connaissances fiables, d'orienter les choix techniques et de limiter l'usage de ressources de calcul.

Le travail rapporté ici constitue une première phase exploratoire et de multiples améliorations peuvent être envisagées : généralisation à d'autres indicateurs, optimisation brique par brique de la pipeline, exploration de modèles de langage propriétaires, traitement des « grandes » collectivités, ... L'approche à suivre dépendra des priorités dites métier, relatives à SISPEA, et des contraintes techniques imposées, notamment en termes de coûts financier et environnemental. Dans tous les cas, il sera nécessaire de mettre en place un outil de suivi des performances de NARVAL car les métriques calculées jusqu'à présent sur un petit jeu de 45 PDFs pourraient se dégrader pour d'autres PDFs. Enfin, nous espérons qu'à terme, NARVAL permettra de soulager le travail de saisie des collectivités et d'enrichir la base de données bancarisée et diffusée par SISPEA. Nous espérons également que ce travail nourrira la réflexion d'autres services de l'administration publique confrontés à des problèmes similaires d'extraction de données structurées à partir de PDFs.

REMERCIEMENTS

Ce travail a bénéficié des moyens de calcul du SSP Cloud [31], des modèles `Flan-T5`[20] de Google et `Llama3-8b-Instruct`[21] de Meta, de la librairie `Transformers`[41] de Hugging Face[22], des librairies `PyPDF2`[14] et `PDFPlumber`[16] pour la lecture des PDFs et de plusieurs autres librairies python standards.

---

[1] Y. Assis *et al.*, TSM **3**, 31 (2021).
[2] « API Scraptablo, » (2023), pour l'extraction de tableaux à partir de PDFs.
[3] « Projet UniBSV, » (2023), pour l'extraction d'informations à partir de Bulletins de Santé des Végétaux (BSV) au format PDF.
[4] « Projet EU Tax Observatory, » (2024), pour l'extraction de tableaux à partir de PDFs. Observatoire européen de la fiscalité et Association Data For Good.
[5] Y. Zou *et al.*, (2023), arXiv:2312.17264.
[6] Q. Zhang *et al.*, (2024), arXiv:2410.21169.
[7] M. Schilling-Wilhelmi *et al.*, (2024), arXiv:2407.16867.
[8] Référentiel général pour l'IA frugale, piloté par l'Ecolab du Commissariat Général au Développement Durable, en partenariat avec l'AFNOR (2024). Voir `https://telechargement.afnor.info/normalisation-afnor-spec-ia-frugale`.
[9] « Les grands défis de l'IA générative, » (2023), livre blanc, piloté par l'association Data For Good.
[10] Programme Confiance.ai. Voir le livre blanc (2022).
[11] Voir `https://www.services.eaufrance.fr/`.
[12] Le modèle de RPQS proposé par SISPEA pour les services d'eau potable est disponible à l'adresse `https://www.services.eaufrance.fr/cms/uploads/RPQS_eau_potable_2022_7fc728871d.pdf`.
[13] En effet, ce projet NARVAL a été mené en utilisant uniquement les ressources du SSP Cloud[31] et un PC portable basique. Nous n'avons eu accès ni à un grand cloud public/privé, ni à une infrastructure interne de calcul avec GPUs.
[14] M. Fenniak *et al.*, « PyPDF2 library, » (2022).
[15] Nous avons également essayé d'utiliser la librairie Python `PDFPlumber`[16] pour effectuer ces extractions. Le texte est en général un peu mieux extrait qu'avec `PyPDF2` mais les tableaux sous forme de texte ont une structure légèrement plus complexe rendant plus difficiles l'interprétation du texte par les SLMs. Pour cette raison, nous avons privilégié `PyPDF2` mais des tests complémentaires seraient souhaitables.
[16] J. Singer-Vine *et al.*, « PDFPlumber library, » (2025).
[17] Dans le PDF problématique, la table des matières s'étale sur deux pages et seule la première est détectée (car la seconde, qui ne contient que deux lignes, n'est pas au format adéquat pour notre détecteur).
[18] Wes McKinney, in *Proceedings of the 9th Python in Science Conference* (2010) pp. 56 – 61.
[19] Nous avons comparé les performances de NARVAL lorsque le `SLM` est interrogé 1) uniquement pour les indicateurs non extraits par le `Table Extractor` et 2) aussi pour les indicateurs extraits par le `Table Extractor` mais à valeurs non-chiffrées (par exemple « Non disponible »). En moyenne, il est préférable de suivre la stratégie 1) qui génère en général plus de faux négatifs mais moins de faux positifs. Toutefois, pour certains PDFs, les performances sont meilleures en adoptant la stratégie 2).
[20] H. W. Chung *et al.*, (2022), arXiv:2210.11416.
[21] A. Grattafiori *et al.*, (2024), arXiv:2407.21783.
[22] Voir `https://huggingface.co/`.
[23] L'un de nos premiers tests a consisté à comparer la métrique d'accuracy (définie en section V A sur un jeu test de 15 PDFs (jeu $\mathcal{D}_1$ défini en section IV) lorsque le modèle `Flan-T5-xl` ou `Llama3-8b-Instruct` est utilisé avec son propre format de prompt, toute autre paramètre égal

par ailleurs. Nos résultats ont alors montré une accuracy de 41% pour le premier et 77% pour le second.

[24] Les noms des « petites » collectivités traitées apparaissent sur la figure 5. Il suffit de saisir leur nom sur le site `https://services.eaufrance.fr/mon-territoire`, de s'assurer de choisir l'échelle de la collectivité et non de la commune puis de choisir l'année pour pouvoir télécharger le rapport RPQS ou RAD correspondant. NARVAL intègre également un module de *scraping* permettant de télécharger en masse plusieurs RPQS ou RAD à partir de SISPEA.

[25] Nous estimons qu'environ 5% des RPQS sur SISPEA sont scannés, du moins pour les "petites" collectivités.

[26] Sauf un, celui de la collectivité Allain pour l'exercise 2021. D'autres RPQS de $\mathcal{D}_{1+2}$ ont toutefois un *template* très similaire à celui des télé-RPQS.

[27] Sauf celui de la commune de Saint-Mihiel, pour l'exercice 2020, qui traite dans un même rapport des services d'eau potable et d'assainissement.

[28] Nous avons comparé les valeurs saisies dans SISPEA aux valeurs indiquées dans les rapports pour les 15 PDFs du jeu $\mathcal{D}_1$. Environ 22% des valeurs d'indicateurs diffèrent. Les raisons sont multiples : valeur SISPEA arrondie par rapport à la valeur PDF (conformément aux règles d'arrondis SISPEA), valeur PDF arrondie par rapport à la valeur SISPEA, valeurs SISPEA et PDF complètement différentes, valeur non inscrite dans le PDF mais saisie sur SISPEA (les collectivités pouvant saisir les valeurs des indicateurs *a posteriori*), valeur non saisie sur SISPEA mais inscrite dans le PDF.

[29] Il se peut en effet que les collectivités ayant fait ce travail de saisie aient rédigé des RPQS plus soignés et plus facilement interprétables par NARVAL que celles n'ayant pas procédé à la saisie.

[30] Afin d'éviter toute confusion du `SLM` entre ces indicateurs, seules les questions du type « Quelle est la valeur de l'indicateur P203.3 en 2021 ? » sont posées, en utilisant le code P203.3 comme mot-clé à l'étape de segmentation (et de même pour P204.3 et P205.3). De plus, nous avons abandonné l'idée d'inclure dans le prompt une instruction demandant au `SLM` de convertir l'information « [non] conforme » ou « case cochée [non] oui » en [0%] 100%.

[31] Voir `https://datalab.sspcloud.fr`.

[32] B. Courty *et al.*, « mlco2/codecarbon: v2.4.1, » (2024).

[33] T. Simon *et al.*, in *HotCarbon'24 - 3rd Workshop on Sustainable Computer Systems* (Santa Cruz, United States, 2024).

[34] Il faudrait alors optimiser l'étape de chargement des poids du modèle sur GPU. Jusqu'à présent, nous avons utilisé comme moyens de calcul le SSP Cloud[31] et télécharger à chaque lancement d'un nouveau service les poids du modèle `LLama3-8b-Instruct` (16 Go) à partir de Hugging Face, ce qui prend 5 à 10 minutes. Il serait sans doute préférable de charger une seule fois les poids du modèle sur un bucket S3 puis de charger les poids sur le GPU à chaque ouverture d'un nouveau service, *a fortiori* pour un modèle plus gros.

[35] L. Martin *et al.*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020) p. 7203.

[36] Peut-être qu'un classifieur à base de forêts aléatoires peut suffire pour identifier les tableaux pertinents et qu'un modèle BERT léger peut permettre d'extraire le nom de la colonne des valeurs.

[37] En pratique, les RPQS pouvant faire plus de 100 pages, il sera sûrement trop coûteux de donner tout le PDF comme contexte au LLM. Il sera donc toujours nécessaire de segmenter le PDF mais on pourra par exemple concaténer toutes les pages identifiées par mots-clés de sorte qu'il ne sera plus nécessaire de sélectionner une réponse parmi plusieurs valeurs candidates.

[38] La librairie Ecologits permet d'estimer le coût environnemental lié à l'utilisation par API d'un modèle d'IA générative. Voir `https://ecologits.ai/latest/`.

[39] Voir `https://www.langchain.com/`.

[40] Dans le cas des variables définies à l'échelle d'une station, il faudra vérifier que le nom des stations dans le référentiel SISPEA (à utiliser dans la question posée au `SLM`) est bien celui employé dans les rapports PDFs. A noter aussi que souvent, les variables intermédiaires ne sont pas mentionnés dans les RPQS ou RAD.

[41] T. Wolf *et al.*, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computational Linguistics, 2020) p. 38.

[42] Voir `https://services.eaufrance.fr/indicateurs` pour une liste complète des indicateurs SISPEA.

Annexe A: Indicateurs SISPEA d'assainissement collectif

La liste des 19 indicateurs SISPEA d'assainissement collectif traités à ce jour par NARVAL [42] est donnée dans la table VI.

Annexe B: Règles de labellisation

Dans certains cas, il n'est pas évident de savoir quelle est la valeur d'un indicateur (servant de référence pour l'évaluation) à partir des informations fournies dans le PDF. Nous nous sommes donc fixés un certain nombre de règles afin de lever les ambiguités. Celles-ci sont listées dans la table VII.

Code	Libellé de l'indicateur	Unité
D201.0	Nombre d'habitants desservis	sans unité
D202.0	Nombre d'autorisations de déversement d'effluents d'établissements industriels	sans unité
D203.0	Quantité de boues issues des ouvrages d'épuration	Tonne de matières sèches (tMS)
D204.0	Prix TTC du service au m3	€/m3
P201.1	Taux de desserte par des réseaux de collecte des eaux usées	%
P202.2B	Connaissance et gestion patrimoniale des réseaux de collecte des eaux usées	point
P203.3	Conformité de la collecte des effluents aux prescriptions nationales issues de la directive ERU	%
P204.3	Conformité des équipements d'épuration aux prescriptions nationales issues de la directive ERU	%
P205.3	Conformité de la performance des ouvrages d'épuration du service aux prescriptions nationales issues de la directive ERU	%
P206.3	Boues évacuées selon des filières conformes	%
P207.0	Montant des actions de solidarité	€/m3
P251.1	Débordements d'effluents chez les usagers	taux pour 1000 habitants
P252.2	Points de curage fréquent du réseau	taux pour 100 km
P253.2	Renouvellement des réseaux de collecte des eaux usées	%
P254.3	Conformité des performances des équipements d'épuration au regard des prescriptions de l'acte individuel	%
P255.3	Connaissance des rejets au milieu naturel	point
P256.2	Durée d'extinction de la dette de la collectivité	année
P257.0	Taux d'impayés sur les factures d'eau	%
P258.1	Taux de réclamations	taux pour 1000 abonnés

TABLE VI. Liste des 19 indicateurs SISPEA d'assainissement collectif, avec leurs unités.

Indicateurs	Valeur ou formulation renseignée dans le PDF	Valeur extraite
P203.3, P204.3, P205.3	« conforme » ou case cochée « oui »	100%
	« non conforme » ou case cochée « non »	0%
P254.3	« conforme »	100%
	« non conforme »	`NULL`
P256.2	Encours de la dette explicitement indiqué comme étant de 0€	0 année
	Dette indiquée comme remboursée en l'année $N + X$, N étant l'année d'exercice	X années
	Informations financières sur l'état de la dette sans que la durée d'extinction de la dette soit explicitement renseignée [a]	`NULL`
D203.0	Quantités de boues indiquées par station d'épuration, sans mention explicite du total pour l'ensemble des stations[b]	Somme des quantités de boues
D204.0	Prix TTC en € mentionné pour 120m^3 mais pas en €/m^3	Prix en €/m^3 après division par 120
	Détails sur le prix du service sans que le prix pour 120m^3 soit explicitement mentionné [a]	`NULL`
Autres	Valeur non renseignée explicitement dans l'unité attendue	`NULL`
Tous	« NEANT »	`NULL`
Tous	Valeur non renseignée pour l'année d'exercice N mais pour une autre année [c]	`NULL`
Tous	Contradiction entre le(s) tableau(x) et le texte, avec valeur du tableau incohérente (coquille, oubli de convertir le prix en €/m^3, ...)	Valeur du texte
	Contradiction entre le(s) tableau(x) et le texte, avec valeur du tableau cohérente	Valeur du tableau

[a] Dans certains cas, il n'est pas exclu qu'un annotateur expert serait capable d'extraire la valeur de l'indicateur à partir des informations fournies dans le PDF.

[b] En réalité, il est attendu dans SISPEA de saisir la quantité de boues par station d'épuration mais ce niveau de détail n'est pas traité dans la version actuelle de NARVAL.

[c] Notamment pour D204.0, la valeur de l'année N est définie comme celle au 1er janvier de l'année $N + 1$.

TABLE VII. Ensemble de règles utilisées pour labelliser les 45 PDFs de $\mathcal{D}_{1+2}$.