Qualité des données dans le processus d'ingestion pour les grands modèles de langage : pratiques et défis

Katy Fokou

Smals research

Smals
ICT for society

"AI in theory should lead to greater fairness, everyone is judged according to the same rules"

Les promesses de l'IA

AlphaGo marked the birth of modern AI. This is
the world chang

By technology reporter James Purtill f

ABC Science | Science and Tech

Tue 24 Oct 2023

New AI
cancer,
predict

Eka

Lee Sedol analysing the game during the series with AlphaGo in 2016. *(Getty: Handout)*

How OpenAI's ChatGPT has changed the world in just a year

The generative AI chatbot has helped kickstart a multibillion dollar industry.

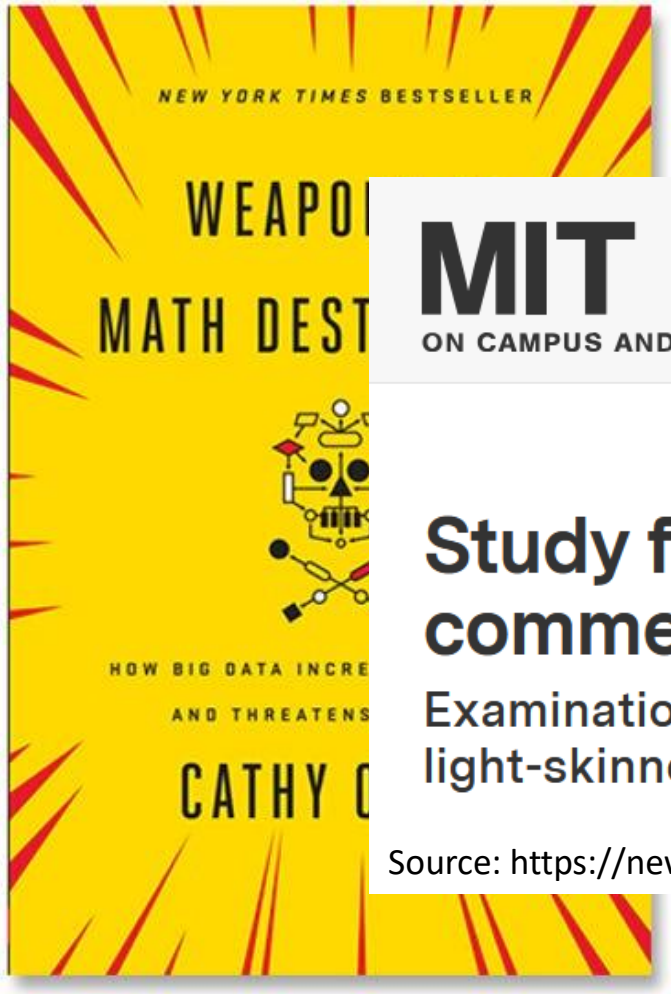
Tarant
ola

Andrew Tarantola
Former Senior Editor
Thu, Nov 30, 2023 · 11 min read

3

Cependant…

MIND THE DATA



MIT News
ON CAMPUS AND AROUND THE WORLD

SUBSCRIBE

Study finds gender and commercial artificial-int

Examination of facial-analysis softw
light-skinned men, 34.7 percent for

Source: https://news.mit.edu/2018/study-finds-gender-

De toeslagenaffaire toont aan dat we uitlegbare AI-regels nodig hebben

UvA-onderzoeker Błażej Kuźniacki dringt aan op meer transparantie rond AI

13 februari 2023

Duizenden ouders werden door de Nederlandse belastingdienst ten onrechte beschuldigd van fraude door discriminerende algoritmes. De gevolgen voor gezinnen waren desastreus. Maar het feit dat het schandaal uiteindelijk aan het licht kwam, kan erop

Importance des données dans l'IA

- L'IA apprend à partir de données, le développement de modèles de *machine learning* nécessite une **large quantité de données** de qualité.
- Les données sont le **moteur vital** de l'IA.
- "**Garbage in, garbage out**": les données influencent la qualité du résultat d'un système d'IA.

> *"Instead of focusing on the code, companies should focus on developing systematic engineering practices for improving data in ways that are reliable, efficient, and systematic. In other words,* ***companies need to move from a model-centric approach to a data-centric approach.****"*
>
> — Andrew Ng, CEO and Founder of LandingAI

Model-centric AI
Focus sur l'architecture du modèle pour améliorer les performances du modèle.

→

Data-centric AI
Focus sur la qualité des données pour améliorer les performances du modèle.

Smals ICT for society

Prérequis pour des données “AI-ready”:

- Une infrastructure pour collecter et stocker les données
- Des outils qui permettent de développer rapidement un pipeline
- L’annotation des données par des experts du domaine (chronophage)
- Un système de contrôle qualité et validation
- Un système de gouvernance et monitoring
- Garantir la protection et la sécurité des données
- **60 to 80 % du temps de développement de modèles est dédié au traitement de données.**

> *“Data readiness for AI is not something you can build once and for all nor that you can build ahead of time for all your data. It is a process and a practice based on availability of metadata to align, qualify and govern the data.”*
>
> What Makes Data AI-Ready?, Gartner

Smals ICT for society

Les défis des données dans la pratique

- **Technique**
  - Difficile de collecter les données, manque d'infrastructure.
  - Manque de gouvernance.
- **Organisationnel**
  - Données dispersées et pas faciles d'accès.
  - Fonctionnement en silo.
  - Difficulté d'échange de données entre institutions.
- **Légal**
  - Pas de lignes directrices claire quand à ce qu'on peut faire et ne pas faire avec les données.
- **Qualité**
  - Non représentatives et pas assez diversifiées, pourrait introduire des biais.
  - Peu ou mal labellisées.
- **Ressource**
  - Données complexes. Besoin d'experts du domaine pour déterminer quelles données sont utiles pour le modèle et comment les interpréter.

Smals ICT for society

Focus sur l’IA générative et les grands modèles de langage

Smals
ICT for society

Qu’est que l’IA générative?

- Sous-domaine de l’IA.
- Plus un modèle est « à usage général », plus il nécessite de la puissance de calcul et des €€€.
- **Modèles d’IA qui génèrent:**

Résumés, e-mails, rapports, documentation, idées.

Images à partir de descriptions.

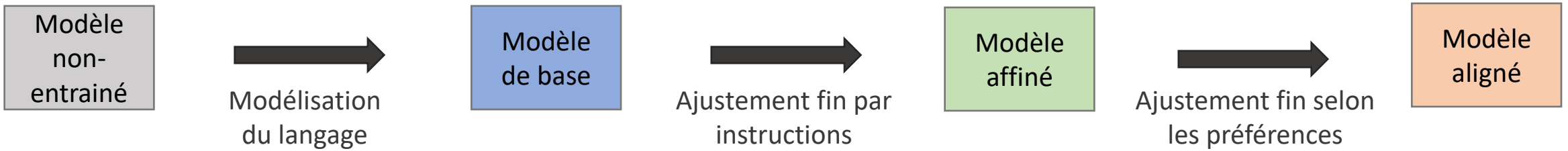
Voix humaine synthétique.

Animations, montage vidéo.

Code logiciel.

Smals ICT for society

Entraînement des grands modèles de langage (LLM)



Modélisation du langage

- Le modèle apprend à prédire le prochain **token** dans une séquence, la génération de contenu est donc un processus probabiliste.
- Entraînement non supervisé sur un grand jeu de données.

Ajustement fin par instructions

- Le modèle de base est entraîné à suivre les instructions humaines de manière plus efficace.
- Entraînement supervisé.

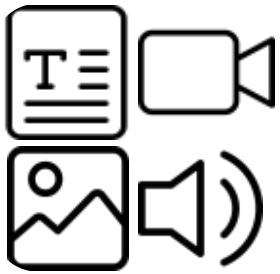
Ajustement fin selon les préférences

- Le modèle de base est entraîné à suivre les instructions humaines de manière plus efficace.
- Reinforcement Learning from Human Feedback

Smals ICT for society

Données d’entraînement pour les LLM

- **Principalement des données publiques**
- **Sources:**
  - Web, media sociaux, principalement Wikipedia, Reddit, X, …
  - Livres (Projet Gutenberg)
  - Bulletins d’actualités
  - Publications de recherche (Pubmed, Researchgate, arXiv)
  - Code (Github, stack exchange)
  - Données synthétiques
  - …
- **Non structurées, différents formats**



Smals ICT for society

Données d’entraînement pour les LLM

- **Caractéristiques de données de qualité**:
  - Equilibrées, couvrent un large éventail de sujets et divers styles d'écriture et de contextes.
  - Propres, sans erreur de formatage, sans texte non pertinent.
  - Factuelles et précises
  - Récentes
  - Respectent la vie privée et les droits d'auteur.
  - Cohérentes dans le formatage et les valeurs qu'elles représentent.
  - Uniques
  - Dont la provenance et la lignée sont connues

Smals ICT for society

Problème de qualité de données - biais

- Image générée pour “Lawyer” (DALL-E 2)

- Image générée pour “Flight attendant” (DALL-E 2)

Source: Sigal Samuel, “A new AI draws delightful and not-so-delightful images”, Vox, 22/04/2022

Problème de qualité de données – données sensibles

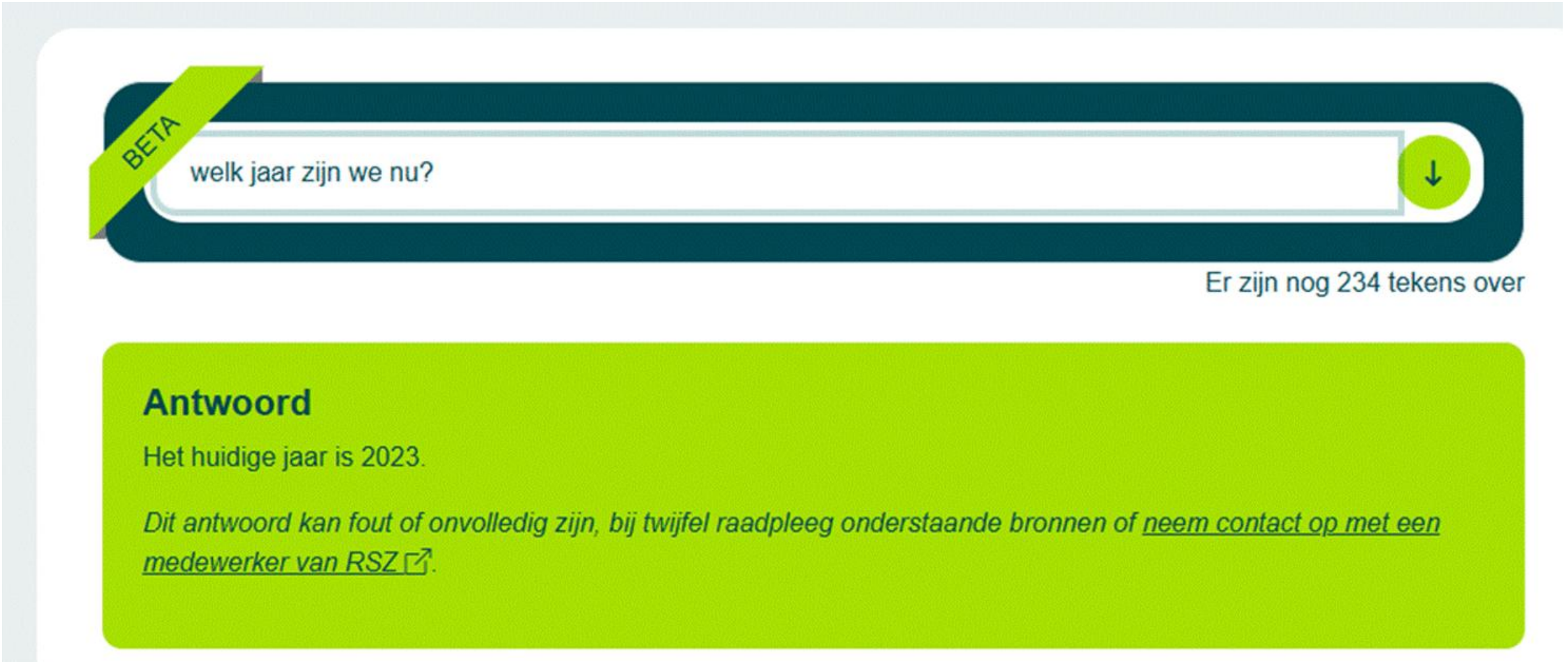
Your chat inputs can be used to train and improve models

«près de la moitié des utilisateurs (42%) partagent des données sensibles de leur entreprise »
(https://www.rtbf.be/article/chatgpt-pres-de-50-des-employes-belges-partagent-trop-de-donnees-professionnelles-sensibles-11232702)

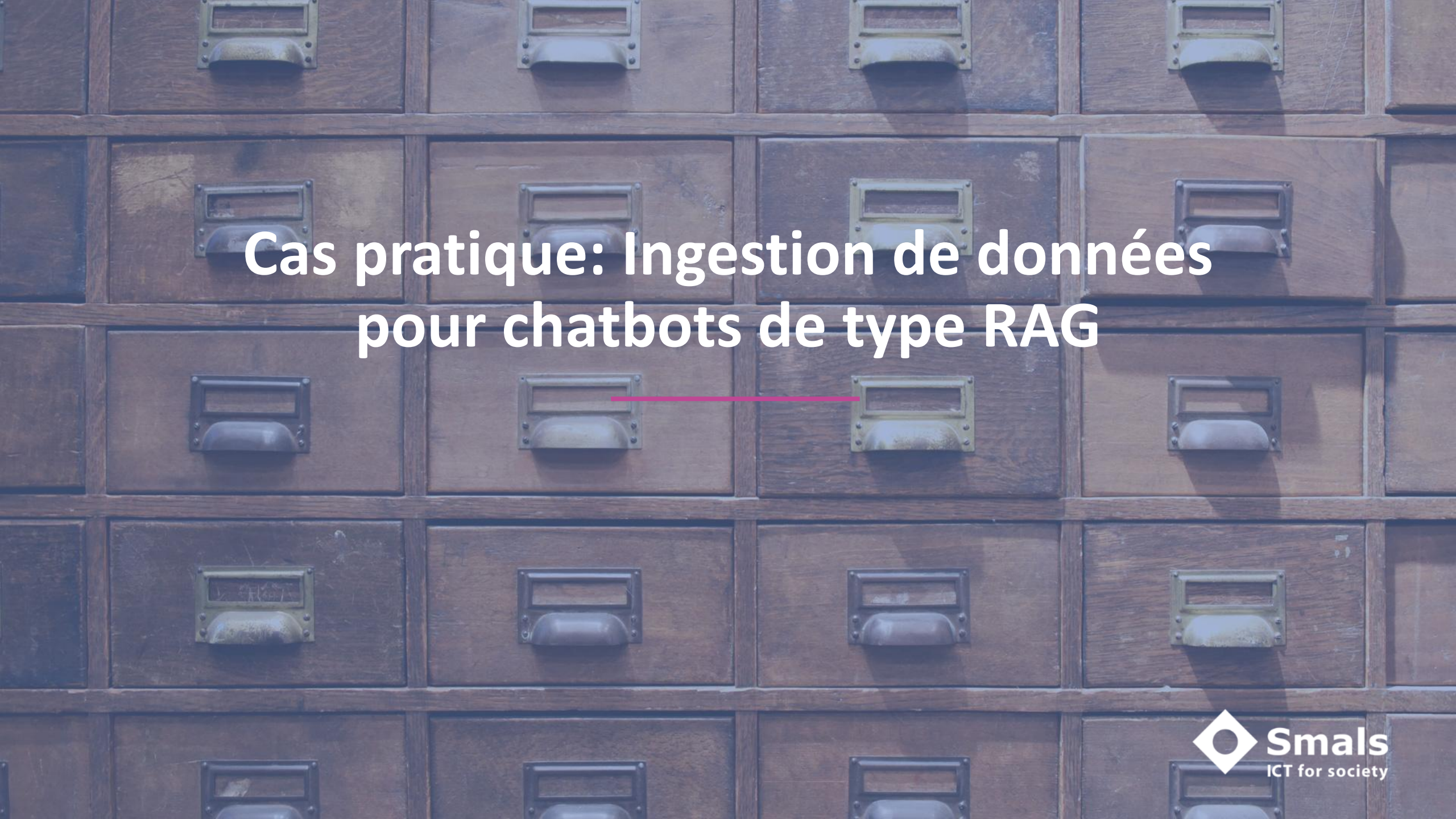
https://www.wired.com/story/chatgpt-poem-forever-security-roundup/

Smals ICT for society

Problème de qualité de données – données non actuelles



Smals ICT for society



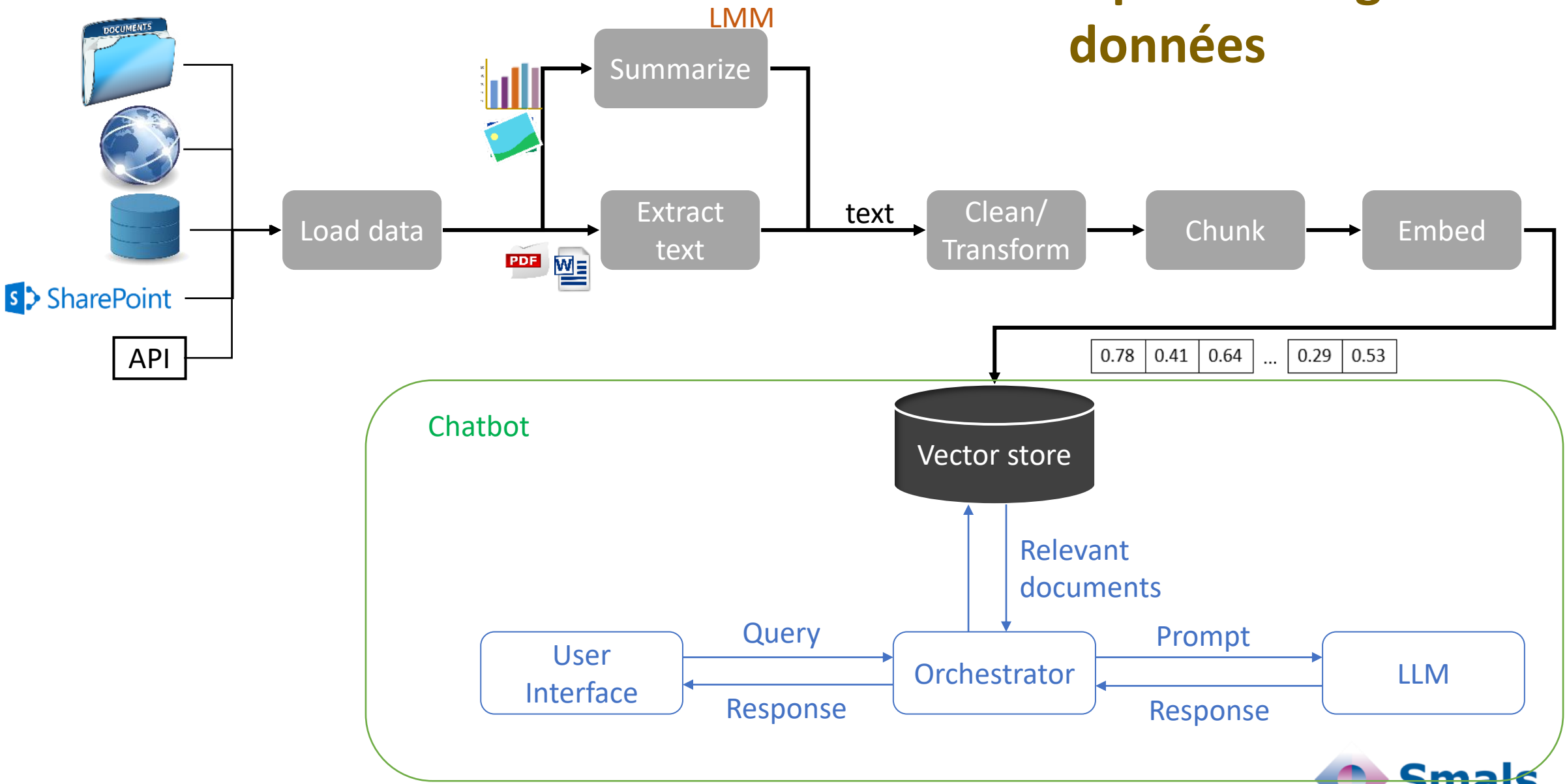
Cas pratique: Ingestion de données pour chatbots de type RAG

Smals
ICT for society

Définition ingestion de données

- **L'ingestion de données, c'est :**
  - Un processus de collecte de données
  - À partir de multiples sources
  - Dans une base de données centralisée
- **L'ingestion de données s'applique à toutes données et peut être très complexe.**
- **Spécificités des données pour des applications basées sur des LLM:**
  - Données non structurées (texte, audio, vidéo)
  - Beaucoup de données proviennent du web
  - De nombreux cas d'usage identifiés pour le RAG ou la recherche sémantique nécessitent la collecte de données provenant de multiples sources et formats :
    - ServiceNow
    - Confluence
    - Web
    - Fichiers Excel, Word, PDF
    - Transcriptions de MS Teams
    - …

Pipeline d'ingestion de données



Smals ICT for society

Ingestion de données - pipeline

- **Charger les données :**
  - Se connecter aux sources
  - Accéder aux données appropriées
- **Extraction de contenu :**
  - Contenu extrait = texte + métadonnées
  - Beaucoup de formats à gérer pour l'extraction → besoin d'un parseur pour chaque format (parseur HTML, parseur PDF, OCR, langage naturel vers requête SQL, …)
- **Nettoyage :**
  - Supprimer caractères spéciaux, espaces, ponctuations, …
  - Supprimer les bannières de cookies, menus issus des données web
  - Supprimer les doublons
  - Anonymiser
- **Enrichir le contenu :**
  - Résumer les images de tableaux
  - Ajouter des références à utiliser par le RAG
- **Découpage (chunking) :**
  - Découpage par taille
  - Découpage par section
  - Découpage sémantique
- **Vectoriser le contenu et l'indexer dans une base de données vectorielle:**
  - Indexation incrémentale ou complète
  - Version

Smals ICT for society

Ingestion de données – principes de gouvernance

- **Ingestion en batch ou en streaming**
- **Gérer les mises à jour du contenu :**
  - Suivre les sources de contenu (à afficher avec la réponse pour les RAGs)
  - Suivre la version et les modifications à la source
  - Comment savoir quelle information est la plus récente ?
- **Qualité des données :**
  - S'assurer que les sources sont fiables
  - Vérifier les doublons, contrôle sémantique !
  - Supprimer le contenu non informatif
  - Vérifier les informations contradictoires
  - …
- **Sécurité :**
  - Sécuriser le transfert des données, éviter les altérations (notamment via le web !)
  - Gérer les droits d'accès lors de l'insertion du contenu dans la base de données vectorielle (transférer les droits d'accès)
- **Monitoring**

Smals ICT for society

Ingestion de données – recommandations

- **Commencer par le profilage des données:**
  - Le profilage des données est le processus de revue, d'analyse et de synthèse d'un ensemble de données afin de comprendre sa structure et de corriger les éventuels problèmes.
  - Permet de comprendre et d'évaluer la qualité des données.
  - Étapes :
    - Comprendre le format, la structure, le type de données, la distribution
    - Comprendre le contenu, le contexte, identifier les incohérences
    - Identifier les relations entre les sets de données (s'ils proviennent de sources différentes)
- **Utiliser les LLM pour préparer et nettoyer les données :**
  - Détection des données personnelles (PII), remplacement des PII par des tirets ou des noms synthétiques.
  - Vérification grammaticale et correction.
  - Classification des données : retenir les sujets pertinents, enlever le contenu nuisible.
- **Utiliser un LLM pour générer des données synthétiques.**

Smals ICT for society

Cas pratiques: Les défis rencontrés

Smals
ICT for society

Défis de l’ingestion de données

Extraction données web

- **Websearch RAG:**
  - Données ingérées « à la volée »
  - Difficile de garantir la fiabilité des données
  - À utiliser avec des sources de confiance
- **Accéder au bon site web :**
  - Des outils existent pour interroger le web
  - Trop de résultats, besoin de filtrage
- **Scraping:**
  - Utiliser un outil adapté (exécution JavaScript)
  - Captcha, bloqueurs de robots
  - Extraire uniquement les informations utiles du HTML
- **Droits d’accès :**
  - Connexion requise

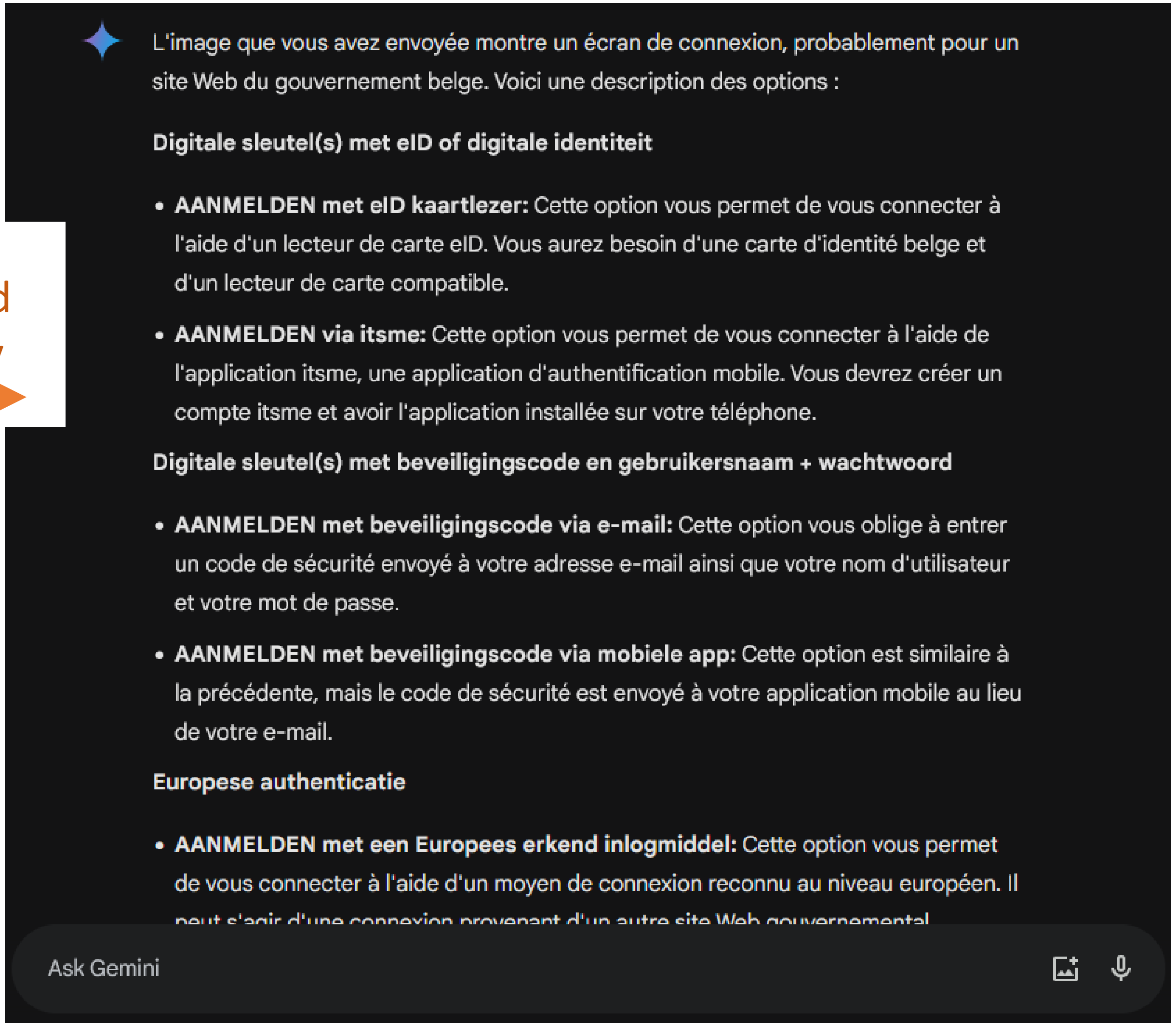
Extraction PDF

- **Semble structuré mais:**
  - Mélange de texte, d’images et de tableaux
  - L’OCR ne capture pas la mise en page
- **Extraction et découpage logique :**
  - Utile pour les documents longs
  - Découpage par section
  - Difficulté de bien délimiter les sections

Smals ICT for society

Ingestion d'images

LMM-generated summary



OCR Extraction

```
ELEMENT #60
        - Category: Image
        - Text: Kies uw digitale sleutel om aan te melden Hulp nodig? Digitale sleutel(s) met elD of digitale identiteit & AANMELDEN AANMELDEN T met elD kaartleze
r via itsme Je itsme-account aanmaken Digitale sleutel(s) met beveiligingscode en gebruikersnaam + wachtwoord AANMELDEN AANMELDEN met beveiligingscode via e-mail m
et beveiligingscode via mobiele app Europese authenticatie lam AANMELDEN © met een Europees erkend inlogmiddel
```

Ingestion d’images: open-source vs commercial

Image: Itsme Aanmelden

Llava 13b

The image appears to be a screenshot of a digital identity verification process from a government or public service portal. Here is the information summarized:

1. **Title**: "Digital identification for people who are not registered in the Netherlands."
2. **Instructions**: There is a note indicating that the user should only use their email address to set up digital identification, and not to share personal details, such as a passport or ID card number.
3. **Options**: The image shows two options for creating digital identification: "Using a code" and "With a Dutch driving licence." There are also symbols indicating options like using an email address, a mobile phone, and other methods of verification.
4. **Logos**: There are logos for "Aanmelden AAN MEEDEM VOLGEN" which translates to "Sign up with us now," suggesting the user can sign up for digital identification services. Additionally, there is a logo representing Europe and possibly an indication of European Union support or involvement in this process.
5. **Text**: There are phrases such as "Digital identification with your email address" and "Digital identification with your Dutch driving licence," which describe the two methods offered for creating digital ID.
6. **Buttons/Links**: The image includes buttons or links related to starting a new sign-up process, accessing different services, and managing existing accounts. These elements suggest an interactive user interface typical of online registration or account management systems.

GEMINI

L'image que vous avez envoyée montre un écran de connexion, probablement pour un site Web du gouvernement belge. Voici une description des options :

Digitale sleutel(s) met eID of digitale identiteit

- **AANMELDEN met eID kaartlezer:** Cette option vous permet de vous connecter à l'aide d'un lecteur de carte eID. Vous aurez besoin d'une carte d'identité belge et d'un lecteur de carte compatible.
- **AANMELDEN via itsme:** Cette option vous permet de vous connecter à l'aide de l'application itsme, une application d'authentification mobile. Vous devrez créer un compte itsme et avoir l'application installée sur votre téléphone.

Digitale sleutel(s) met beveiligingscode en gebruikersnaam + wachtwoord

- **AANMELDEN met beveiligingscode via e-mail:** Cette option vous oblige à entrer un code de sécurité envoyé à votre adresse e-mail ainsi que votre nom d'utilisateur et votre mot de passe.
- **AANMELDEN met beveiligingscode via mobiele app:** Cette option est similaire à la précédente, mais le code de sécurité est envoyé à votre application mobile au lieu de votre e-mail.

Europese authenticatie

- **AANMELDEN met een Europees erkend inlogmiddel:** Cette option vous permet de vous connecter à l'aide d'un moyen de connexion reconnu au niveau européen. Il peut s'agir d'une connexion provenant d'un autre site Web gouvernemental

Ask Gemini

Extraction de texte à partir d'un PDF (Docling)

Extract elements

```
Arrêt
 (Type: section_header - Page: 1 - Closest heading: )

Contradictoire Définitif
 (Type: text - Page: 1 - Closest heading: Arrêt)

*  Sécurité  sociale -travailleurs  salariés -réduction  groupe-cible « premiers  engageme
de l ' effectif au moment  de l ' engagement -loi-programme (I) du 24 décembre 2002 (art.
 (Type: list_item - Page: 1 - Closest heading: Arrêt)

N° d'ordre
 (Type: page_header - Page: 1 - Closest heading: )

Cour du travail de Liège, division Liège -2020/AL/317 -p. 2
 (Type: page_header - Page: 2 - Closest heading: )

N° d 'ordre
 (Type: page_header - Page: 2 - Closest heading: )

EN CAUSE :
 (Type: section_header - Page: 2 - Closest heading: )

L ' OFFICE NATIONAL DE SECURITE SOCIALE , dont les bureaux sont situés à 1060 BRUXELLES, p
treprises  sous  le  numéro 0206.731.645,
 (Type: text - Page: 2 - Closest heading: EN CAUSE :)

ayant comparu pour conseil Maître Dominique DRION, avocat à 4000 LIEGE, rue Hullos 103105,
 (Type: text - Page: 2 - Closest heading: EN CAUSE :)

partie  appelante  au  principal,  intimée  au  principal  et  intimée  sur  incident,  c
 (Type: text - Page: 2 - Closest heading: EN CAUSE :)

CONTRE :
 (Type: section_header - Page: 2 - Closest heading: )

1. Maître J.P. R. , avocat, dont le cabinet est situé à 4020 LIEGE, quai des Ardennes 65,
ce,
```

text

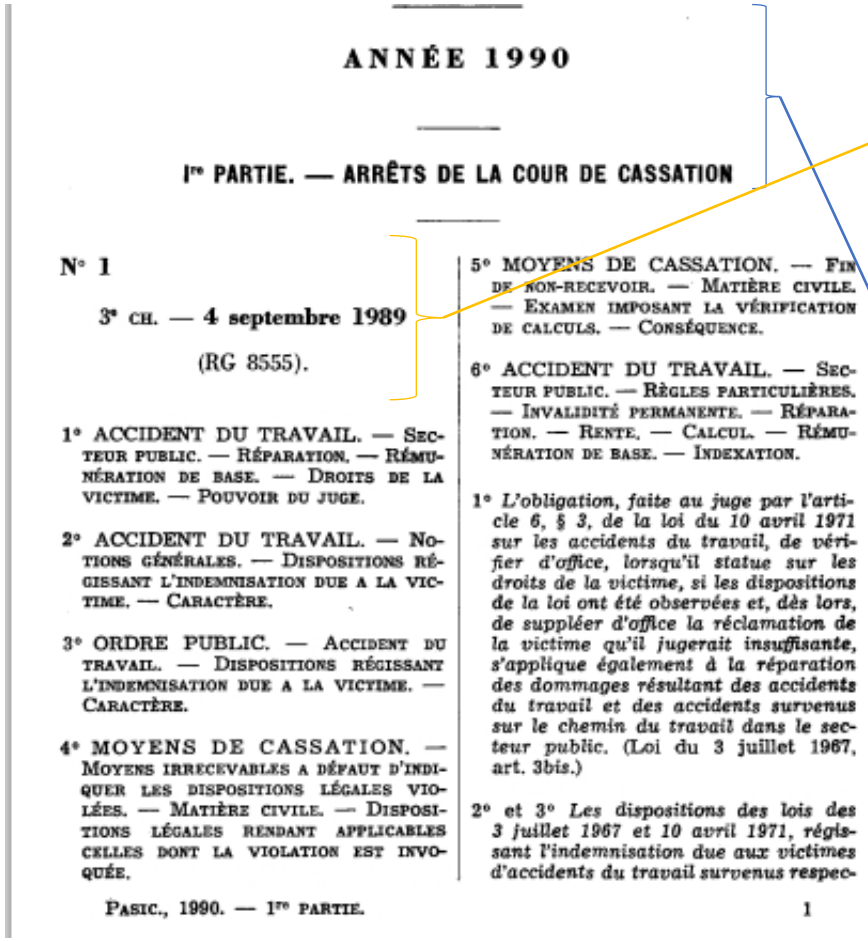
metadata

Extract in markdown format

```
## Arrêt

Contradictoire Définitif

- *  Sécurité  sociale -travailleurs  salariés -réduction  groupe-cible « premiers  engagements » -u
ul de l ' effectif au moment  de l ' engagement -loi-programme (I) du 24 décembre 2002 (art. 342 à 3

## EN CAUSE :

L ' OFFICE NATIONAL DE SECURITE SOCIALE , dont les bureaux sont situés à 1060 BRUXELLES, place  Vict
treprises  sous  le  numéro 0206.731.645,

ayant comparu pour conseil Maître Dominique DRION, avocat à 4000 LIEGE, rue Hullos 103105, et ayant

partie  appelante  au  principal,  intimée  au  principal  et  intimée  sur  incident,  ci-après dé

## CONTRE :

- 1. Maître J.P. R. , avocat, dont le cabinet est situé à 4020 LIEGE, quai des Ardennes 65, ayant co
ence,
- 2. Maître A.B. , avocat, dont le cabinet est situé à 4000 LIEGE, quai de Rome 19/8, n ' ayant pas

agissant en leur qualité de curateurs à la faillite de la sprl MD, ci-après dénommée « la SPRL MD »,

parties  intimées  au  principal,  appelantes  au  principal  et  appelantes  sur  incident,  ci-apr
» ou encore « la SPRL MD ».

.

.      .

## I. INDICATIONS DE PROCEDURE

- 1. La Cour a tenu compte des pièces figurant en forme régulière dans le dossier de la procédure à
- -le  premier  jugement  attaqué,  rendu  contradictoirement  entre  les  parties  le  18 novembre
6 e Chambre (R.G. 18/3388/A)

- -le  second  jugement  attaqué,  rendu  contradictoirement  entre  les  parties  le  11  mai 2020
```


ICT for society

Extraction de texte à partir d'un PDF (Docling)



ANNÉE 1990

Ire PARTIE. — ARRÊTS DE LA COUR DE CASSATION

N° 1

3e CH. — 4 septembre 1989

(RG 8555).

1° ACCIDENT DU TRAVAIL. — SECTEUR PUBLIC. — RÉPARATION. — RÉMUNÉRATION DE BASE. — DROITS DE LA VICTIME. — POUVOIR DU JUGE.

2° ACCIDENT DU TRAVAIL. — NOTIONS GÉNÉRALES. — DISPOSITIONS RÉGISSANT L'INDEMNISATION DUE A LA VICTIME. — CARACTÈRE.

3° ORDRE PUBLIC. — ACCIDENT DU TRAVAIL. — DISPOSITIONS RÉGISSANT L'INDEMNISATION DUE A LA VICTIME. — CARACTÈRE.

4° MOYENS DE CASSATION. — MOYENS IRRECEVABLES A DÉFAUT D'INDIQUER LES DISPOSITIONS LÉGALES VIOLÉES. — MATIÈRE CIVILE. — DISPOSITIONS LÉGALES RENDANT APPLICABLES CELLES DONT LA VIOLATION EST INVOQUÉE.

5° MOYENS DE CASSATION. — FIN DE NON-RECEVOIR. — MATIÈRE CIVILE. — EXAMEN IMPOSANT LA VÉRIFICATION DE CALCULS. — CONSÉQUENCE.

6° ACCIDENT DU TRAVAIL. — SECTEUR PUBLIC. — RÈGLES PARTICULIÈRES. — INVALIDITÉ PERMANENTE. — RÉPARATION. — RENTE. — CALCUL. — RÉMUNÉRATION DE BASE. — INDEXATION.

1° *L'obligation, faite au juge par l'article 6, § 3, de la loi du 10 avril 1971 sur les accidents du travail, de vérifier d'office, lorsqu'il statue sur les droits de la victime, si les dispositions de la loi ont été observées et, dès lors, de suppléer d'office la réclamation de la victime qu'il jugerait insuffisante, s'applique également à la réparation des dommages résultant des accidents du travail et des accidents survenus sur le chemin du travail dans le secteur public.* (Loi du 3 juillet 1967, art. 3bis.)

2° et 3° *Les dispositions des lois des 3 juillet 1967 et 10 avril 1971, régissant l'indemnisation due aux victimes d'accidents du travail survenus respec-*

PASIC., 1990. — Ire PARTIE. 1

```
## No 1

3e  CH.  -

4  septembre  1989

(RG  8555).

- 1 ° ACCIDENT DU TRAVAIL. SEC› TEUR PUBLIC. RÉPARATION. R"ÉMU› NÉRATION  DE
- 2°  ACCIDENT  DU  TRAVAIL.  No› TIONS  GÉNÉRALES.  DISPOSITIONS  RÉ› GISSAN
- 3°  ORDRE  PUBLIC.  ACCIDENT  DU TRAVAIL.  DISPOSITIONS  RÉGISSANT L'INDEMN
- 4° MOYENS  DE  CASSATION.  MOYENS IRRECEVABLES  A  DÉFAUT  D'INDI› QUER  LE
LE.  DISPOSI› TIONS LÉGALES RENDANT APPLICABLES CELLES  DONT  LA  VIOLATION

PASIC.,  1990.  1re  PARTIE.

## PASICRISIE  BELGE

RECUEIL GENERAL DE LA JURISPRUDENCE DES  COURS  ET TRIBUNAUX  ET  DU  CONSEIL

## ANNÉE  1990

## 1re  PARTIE.  ARRETS  DE  LA  COUR  DE  CASSATION

- 5°  MOYENS  DE  CASSATION.  FIN DE  NON-RECEVOIR.  MATIRE  CIVILE. -EXAMEN
NCE.
- 6° ACCIDENT  DU  TRAVAIL.  SEC› TEUR PUBLIC. RGLES  PARTICULIRES. -INVALIDI
ÉMU› NÉRATION  DE  BASE. INDEXATION.
- 1 ° L'obligation, faite au juge par l'arti› cle 6, § 3, de la loi  du  10
```

Données “anonymisées”

Bonjour , Monsieur j'aurais voulu que je puisse avoir les cordonnées de mon mari . en même temps que les miennes . J 'attends votre réponse et vous remercie d'avance .Recevez mes sentiments distingués .	Bonjour, Nous avons reçu votre demande reprise ci-dessous. Afin de pouvoir traiter votre demande, merci de reformuler votre question. Cordialement, Bonjour , Monsieur j'aurais voulu que je puisse avoir les cordonnées de mon mari . en même temps que les miennes . J 'attends votre réponse et vous remercie d'avance .Recevez mes sentiments distingués . Mme Jane Doe

Le nom est enlevé du message original

Le nom est présent dans la réponse au message

Smals ICT for society

Recommandations pour la création de contenu de qualité

- Structurer le texte:
  - Les titres doivent être explicites ; éviter d'écrire des paragraphes sous une forme qui pourrait être confondue avec des titres (comme l'usage des majuscules).
  - Le contenu doit être stocké dans un format facile à parser (XML, JSON, …).
- Taguer les contenus (HTML, XML, images)
- Ajouter des métadonnées pertinentes.
- Supprimer les doublons ; les informations sur un même sujet doivent être regroupées dans une même section.
- Éviter les données sensibles.
- S'assurer que le contenu est cohérent, sans erreur ni ambiguïté.
- Appliquer une gestion de contenu appropriée (gestion des versions, mises à jour, …).
- Rendre le document nativement numérique, éviter les contenus scannés.

Smals ICT for society

Outils utiles

- Airbyte. Moteur d'intégration des données structurées et non structurées utilisé pour alimenter les *data warehouse* ou *data lake*.
- **Unstructured. Outil de construction de pipelines de données pour les LLM. Multiple connecteurs et parsers.**
- **Docling Pour l'extraction de contenu PDFs.**
- Langchain. Outil très versatile de développement d'applications basées sur l'IA générative, s'intègre avec de nombreuses sources de données ainsi que la plateforme Unstructured, propose de nombreuses fonctions de *chunking*.
- LlamaIndex. Outil de développement d'applications basées sur l'IA générative pour la recherche dans les bases de connaissance. LlamaIndex dispose également d'un service de gestion des pipelines de données LLamaCloud.
- Ray. Bibliothèque Python pour la gestion des processus computationnels distribués.
- Récupération des données d'une page web : BeautifulSoup (bibliothèque Python), Playwright, FireCrawl.
- Presidio : Filtrage des données personnelles identifiables.
- **Data quality:**
  - Data Quality Tools & Solutions | IBM
  - Great expectations
  - Whylogs: data profiling and logging
  - Talend Data Quality

Smals ICT for society

Questions

Smals
ICT for society