Wikit

Optimiser ses documents pour les chatbots

Bonnes pratiques et conseils

Wikit

SOMMAIRE

1- Pourquoi ce document ? 2

2- Résumé des bonnes pratiques 2

3- Comment améliorer la compréhension de mon PDF par le chatbot ? 3

1. Le format 3
2. La table des matières 3
3. Les tableaux 4
   a. Exemples de bons tableaux 5
   b. Exemples de tableaux plus difficiles à interpréter 5
4. Les images 6

4- Pour aller plus loin : 7

1. Pourquoi le format PDF est-il difficile à traiter ? 7
2. J'ai des documents PDF que je ne peux pas modifier. Est-ce grave ? 7

Wikit

1- Pourquoi ce document ?

Tout comme les humains, les chatbots sont désormais capables d'extraire de l'information des documents de différentes nature : PDF, Word (.docx), Markdown etc…

Et, tout comme les humains, leur compréhension des documents et de leur contenu est grandement facilitée lorsque le document est bien structuré et que l'information qu'il contient est claire.

Ce document a pour but de fournir quelques conseils et bonnes pratiques de mise en forme de documents de sorte à **optimiser leur compréhension du document** et ainsi **améliorer les réponses qu'ils fournissent**.

2- Résumé des bonnes pratiques

Cette section reprend les différents points à considérer pour faciliter la compréhension d'un document par les chatbots.

- **Éviter les documents type Powerpoint :** ils contiennent souvent beaucoup de schémas, peu d'explications et le format paysage rend l'interprétation de l'ordre dans lequel l'information doit être lue n'est pas toujours clair.
- **Ajouter des éléments de structure :** titres, sous-titres, accompagnés **d'une table des matière** générée automatiquement.
- **Respecter les conventions de formatage :** taille de la police qui dépend du niveau des titres des parties, une seule taille de police pour le corps du texte, etc.
- **S'assurer que les tableaux soit clairs :** la présence de lignes entre les cellules aide beaucoup.
- **Accompagner les images/screenshots d'une description**
- **Avoir une page de garde,** avec le titre en gros.

> *Au final, même pour nous les humains, un document clair, structuré, avec une mise en page sans trop de fioriture est généralement plus agréable à lire et permet d'y récupérer l'information nécessaire plus rapidement.*

Wikit

3- Comment améliorer la compréhension de mon PDF par le chatbot ?

Pour améliorer la compréhension du document par le chatbot, il est nécessaire que le chatbot en comprenne la structure : sections, sous-sections, tableaux. **Plus ces éléments sont évidents**, mieux le document peut être traité et **mieux le chatbot peut y récupérer l'information !**

1. Le format

Quelques recommandations:

- **privilégier l'orientation portrait à paysage** (i.e word à powerpoint): les documents au format portrait viennent souvent de logiciels dont le but principal est l'édition de texte (word, google doc, etc…). Ces logiciels forcent un sens de lecture plus intuitif (haut bas et gauche droite). Même les documents sur plusieurs colonnes peuvent être bien compris ! Au contraire, les logiciels comme powerpoint permettent de mettre en forme des documents plus difficiles à interpréter : l'ordre de lecture d'une diapositive peut-être mal compris par le chatbot, et donc les blocs de texte qui y figurent peuvent être lus dans le désordre. En plus le format "powerpoint" encourage l'ajout d'information difficiles à interpréter (graphiques, schémas) et superflues (masque de diapositive, décorations).

- **respecter les conventions de formatage de style :** il est par exemple communément admis que des titres de section plus gros signifient que le titre est plus important, ou qu'un élément indenté appartient à l'élément qui le précède (c.f table des matières).

- **Avoir une page de garde** dont le seul élément textuel est le titre du document, écrit en gros.

2. La table des matières

La table des matières est un élément important qui permet d'indiquer la structure du document. Au-delà de 2 pages, un document est censé être divisé en sections et donc avoir la table des matières. Pour assurer qu'elle soit bien exploitée par le chatbot, mieux vaut s'assurer qu'elle est un formatage "classique":

- **utiliser les fonctionnalités automatiques de création de table des matière** : c'est le meilleur moyen de s'assurer que le format sois bien standardisé, et que les numéros

de pages soient bien à jour. Et c'est plus rapide que d'écrire une table des matières à la main. 😉

- **s'assurer que la table a un format "classique"**:
    - les titres des parties d'un même niveau sont bien alignés
    - les titres des sous-parties sont indentés d'une tabulation par rapport à la partie à laquelle ils appartiennent.
    - le titre et le numéro de page sont relié par une ligne d'un symbole courant (point, tiret)
    - le numéro de page est aligné sur la gauche

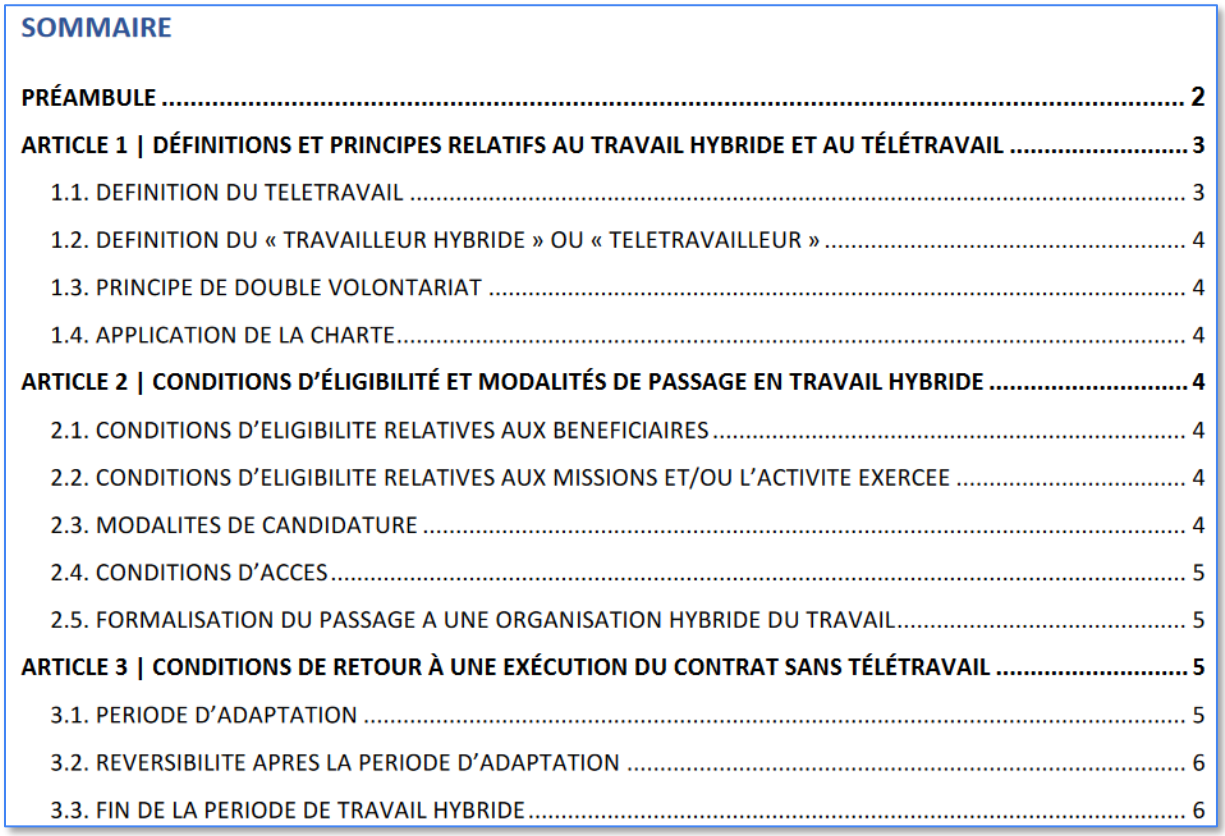


Figure 1 : ☑ Exemple de table des matières idéalement formatée

3. Les tableaux

Les récentes avancées technologiques ont repoussé la capacité des IA à comprendre les tableaux ! Heureusement, car ils sont parfois indispensables pour représenter l'information. Toujours est-il qu'**avoir un tableau avec une structure bien apparente est toujours un plus**.

Voici quelques exemples de bons et moins bon cas.

Wikit

a. Exemples de bons tableaux

	Début des réservations avec paiement immédiat en ligne et en guichet à partir du :	Fin des réservations
Vacances de noël (mardi 26/12/23 au vendredi 05/01/24)	Jeudi 7 décembre 2023 à 12h00 (midi)	Vendredi 15 décembre 2023
Vacances d'hiver (lundi 12/02/24 au vendredi 23/02 /24)	Mardi 23 janvier 2024	Vendredi 02 février 2024
Vacances de printemps (lundi 08/04 au vendredi 19/04)	Mardi 19 mars 2024	Vendredi 29 mars 2024

Figure 2 : ☑ La structure de ce tableau est claire, même si les lignes sont blanches sur fond coloré.

Statut	Textes de référence	Heures d'enseignement	Référentiel ou heures complémentaires possibles ?
Doctorant contractuel Lyon 2 vacataire	Décret 2009-464 du 23 avril 2009	Max 64 HETD[1]	Non
Personnel BIATSS Lyon 2 vacataire	Code Général de la Fonction Publique	Max 64 HETD	
Contractuel LRU L954-3 Lyon 2 vacataire	Article L954-3 du Code de l'éducation	Max 96 HETD	

Figure 3 : ☑ On pourrait penser que la présence de cellules fusionnées pose problème. Elle est toutefois bien prise en charge tant que les lignes sont bien visibles.

b. Exemples de tableaux plus difficiles à interpréter

DIRECTION GÉNÉRALE		
Directrice générale Attachée de direction Secrétaire de direction	<Nom Prénom> <Nom Prénom> <Nom Prénom>	<Numéro> <Numéro> <Numéro>
Directeur délégué au secteur sanitaire Assistante	<Nom Prénom>	<Numéro>

Assets		
Cash and cash equivalents	$ 4,295	$ 4,184
Restricted cash and cash equivalents	631	543
Accounts receivable, net of allowance of $51 and $64, respectively	2,439	2,476
Prepaid expenses and other current assets	1,454	1,462
Total current assets	8,819	8,665
Restricted cash and cash equivalents	2,879	2,865
Investments	11,806	6,247
Equity method investments	800	624
Property and equipment, net	1,853	1,853
Operating lease right-of-use assets	1,388	1,439
Intangible assets, net	2,412	2,269
Goodwill	8,420	8,435
Other assets	397	415
Total assets	$ 38,774	$ 32,812

Figure 4 : ☒ Bien que ces tableaux puissent paraitre clairs, leurs structure n'est pas idéale. Les lignes ne sont que suggérées, les colonnes n'ont pas de nom, la présence de symbole $ uniquement sur certaines lignes est perturbante.

	Centre Val de Loire	Lyon	Rennes	Rouen Normandie	Strasbourg	Toulouse
Arts plastiques études	⌞	⌞	⌝	⌝		
Danse études	⌞	⌞		⌝		⌝
Image études				⌝		
Lumières études			⌝			
Musique études	⌝	⌝	⌝	⌝	⌝**	⌝
Sport études	⌝	⌝	⌝	⌝	⌝	⌝
Théâtre études	⌝	⌞*	⌝	⌝		⌝
Cinéma études		⌞				

⌝ *accessible en 1re année du cursus ingénieur en 5 ans*

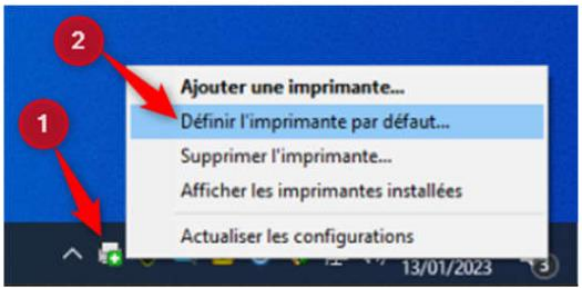
⌞ *accessible en 2e année du cursus ingénieur en 5 ans*

Figure 5 : ☒ La signification des symboles peut être difficile à comprendre. Mieux vaut écrire leur signification dans la cellule.

4. Les images

La technologie progresse, et vite ! Mais **l'interprétation des images par les machines reste un vrai défi,** surtout quand celles-ci sont porteuses d'information difficiles à interpréter comme des graphiques ou des captures d'écran.

Pour s'assurer que leur information ne soit pas perdue, mieux vaut accompagner les images d'une légende qui décrit bien son contenu. Même pour un humain, une légende est souvent indispensable à la compréhension d'un graphique.



Etape 1 : Effectuer un clic droit sur l'icône d'imprimante.
Etape 2 : Sélectionner « imprimante par défaut

Figure 6 : Exemple d'image qui nécessite une annotation, et de son annotation correspondante.

4- Pour aller plus loin :

1. *Pourquoi le format PDF est-il difficile à traiter ?*

Les PDF représentent l'une des principales sources de connaissances qui alimentent les chatbot. Malheureusement ce format de document n'est pas facile à traiter car **il ne contient aucune information sur la structure de son contenu**.

Un fichier classique (.docx, .ppt, etc…) contient de nombreux éléments structurels: blocs de textes, tableaux, tables des matières, numéro de pages, en-tête et pieds de page, etc. Pourtant, **lorsqu'il est converti en .pdf, tous ces éléments sont transformés en un bloc de texte posés sur une page blanche sans aucune information sur la nature de l'élément** : les paragraphes ne sont que des blocs de 1 ligne posés les uns en dessous des autres, les titres ne sont plus que des blocs avec un formatage spécial (couleur, gras, ect…) sans indication de leur niveau, et les tableaux ne sont plus que des petit blocs séparés par des lignes.

Pour un humain cela ne pose aucun problème : nous arrivons à interpréter visuellement la structure du texte, à savoir qu'une partie est une sous-partie d'une autre, etc… Mais pour une machine c'est beaucoup plus compliqué, d'autant que le formatage d'un document peut avoir des formes très (trop) variées !

2. *J'ai des documents PDF que je ne peux pas modifier. Est-ce grave ?*

Non ! 😊

Même si vous ne disposez pas des fichiers originaux qui permettraient d'améliorer le formatage du contenu du PDF, ce n'est pas grave. Ce document ne fourni que des conseils et des **bonnes pratiques pour maximiser les performances d'un chatbot**. **Heureusement**

de bonnes performances peuvent être obtenues même si toutes ces consignes ne sont pas respectées.

Wikit