

Le contexte et les IA génératives

Comprendre, construire et maîtriser le contexte d'une requête

L'IDÉE ESSENTIELLE

Une IA ne travaille pas seulement avec une question : elle travaille avec un **contexte**.

Le contexte correspond à l'ensemble des informations accessibles au modèle au moment où il doit produire une réponse. Il permet d'interpréter la demande, d'identifier l'objectif, de respecter les contraintes et de tenir compte des informations déjà fournies.

Contexte pertinent → moins d'ambiguïté → réponse généralement mieux adaptée

1. COMPRENDRE CE QU'EST LE CONTEXTE

Le contexte donne du sens à une demande. Une phrase isolée peut être ambiguë alors que les informations qui l'entourent permettent de comprendre le sujet, l'objectif ou le niveau de réponse attendu.

- Même question ≠ même réponse : le contexte disponible peut modifier l'interprétation.
- Contexte explicite : consignes, documents, exemples, objectifs et contraintes fournis directement.
- Contexte implicite : informations suggérées ou déduites ; elles sont plus incertaines.

2. LES COMPOSANTS DU CONTEXTE

Un contexte efficace répond à plusieurs questions :

QUI ?	Qui demande ? Pour qui la réponse est-elle destinée ? Quel est le niveau du destinataire ?
POURQUOI ?	Quel est l'objectif réel de la demande ?
QUOI ?	Quelle tâche doit être réalisée ? Quelles informations sont déjà disponibles ?
COMMENT ?	Quel format, quel ton, quelle longueur, quelle structure et quelles contraintes faut-il respecter ?
AVEC QUOI ?	Quels documents, exemples, images, données ou sources peut-on fournir ?
VÉRIFIER ?	Comment contrôler la réponse, les sources et les informations produites ?

3. LE CONTEXTE DANS UNE CONVERSATION

Une conversation permet de construire progressivement le contexte. Chaque nouveau message peut s'appuyer sur les informations déjà échangées.

- Contexte cumulatif : objectifs, corrections, exemples et contraintes s'ajoutent progressivement.
- Les références comme « ceci », « comme avant » ou « conserve cette structure » fonctionnent si l'information correspondante reste accessible.
- Une conversation permet une correction progressive du résultat.
- Attention au risque de dérive : dans une longue conversation, certaines informations anciennes peuvent devenir moins pertinentes.
- Une nouvelle conversation peut disposer d'un contexte différent.

4. LA FENÊTRE DE CONTEXTE

Un modèle ne peut pas nécessairement prendre en compte une quantité infinie d'informations. Sa capacité est limitée par une fenêtre de contexte, généralement exprimée en tokens.

Instructions → Conversation → Documents → Requête → Fenêtre de contexte

Contexte ≠ mémoire permanente. Une fenêtre de contexte contient les informations disponibles pour l'interaction ; elle ne doit pas être confondue avec une éventuelle fonction de mémoire persistante de l'application.

5. LES DOCUMENTS COMME CONTEXTE

Un document peut fournir à l'IA une source d'information spécifique : PDF, document texte, tableur, image, présentation, etc., selon les capacités de l'outil.

Il peut servir à analyser, résumer, transformer ou questionner son contenu.

Dans un système de type RAG, des informations pertinentes sont récupérées dans une source documentaire puis fournies au modèle comme contexte avant la génération.^a

Un document fourni à une IA n'est pas automatiquement une vérité : son contenu doit lui aussi être vérifié.

^a. Un système de type RAG (pour Retrieval-Augmented Generation, ou Génération Augmentée par Récupération) est une architecture d'intelligence artificielle qui combine un système de recherche documentaire avec un modèle de langage (LLM).

6. PLUS DE CONTEXTE N'EST PAS TOUJOURS MIEUX

Un contexte très long peut être contre-productif.

Contexte utile

- Informations pertinentes
- Objectif clairement défini
- Contraintes explicites
- Informations cohérentes
- Sources pertinentes

Contexte problématique

- Informations
- Objectif ambigu
- Contraintes absentes
- Informations contradictoires
- Accumulation de documents

La bonne question n'est donc pas :

« Comment donner le plus d'informations possible ? »

mais :

« Quelles informations sont réellement utiles pour réaliser cette tâche ? »

7. LES ERREURS FRÉQUENTES

- **Demande trop vague** : « Fais-moi un article. » L'IA doit deviner le sujet, le public, le ton et la longueur.
- **Informations essentielles absentes** : ce qui semble évident pour l'utilisateur peut ne pas être disponible pour l'IA.
- **Trop d'informations inutiles** : un contexte volumineux n'est pas nécessairement efficace.
- **Informations contradictoires** : plusieurs règles incompatibles rendent l'interprétation plus difficile.
- **Changement brutal de sujet** : l'ancien contexte peut continuer à influencer la conversation.

8. UNE MÉTHODE SIMPLE POUR CONSTRUIRE SON CONTEXTE

Avant de demander quelque chose à une IA, il est très utile de structurer sa pensée en appliquant, par exemple, la méthode QQOCP, complétée par des critères de validation et des exemples :

QUI ? QUOI ? OÙ ? QUAND ? POURQUOI ?

Puis préciser les modalités pratiques :

COMMENT ? (Méthode, ton, format) **AVEC QUOI ?** (Contraintes, outils, sources)

Pour finir, assurez-vous de définir clairement :

QUEL RÉSULTAT ATTENDU ? **COMMENT LE VÉRIFIER ?**

Lorsque cela est utile, donner un exemple du résultat souhaité (technique du *Few-Shot Prompting*) réduit fortement l'ambiguïté et guide l'IA vers la précision attendue.

Ne demandez pas seulement à l'IA ce qu'elle doit faire. Expliquez-lui dans quel contexte elle doit le faire.

Contexte pertinent → meilleure compréhension de la situation → réponse généralement mieux adaptée.

Le contexte peut être constitué de la requête, des instructions, de la conversation, de documents, de données, d'exemples et des contraintes. Il est limité par la fenêtre de contexte du système.