

FICHE RÉCAP — SÉANCE 1

Sous le capot : tokens, contexte, mémoire et modèles

L'essentiel à retenir, à garder près de son clavier.

1. Les 4 composantes d'un agent IA

COMPOSANTE	C'EST QUOI	EXEMPLES
 Cerveau	Le modèle qui raisonne et rédige	GPT, Claude, Gemini
 Mémoire	Ce qu'il sait de vous et de vos dossiers	Instructions, Projets, vos documents
 Bras	Ce qu'il peut faire dans vos outils	Connecteurs, API, MCP
 Autonomie	Ce qu'il fait sans qu'on le demande	Automatisations, tâches planifiées

2. Tokens — l'unité de mesure de l'IA

- L'IA ne lit pas des mots : elle découpe tout en **tokens** (morceaux de mots). En français, 1 token \approx $\frac{1}{4}$ de mot.
- Ordres de grandeur : un courriel \approx 300 tokens · une page \approx 500 tokens · un rapport de 30 pages \approx 15 000 tokens.
- Tout se paie et se limite en tokens : la longueur de vos questions, des réponses, et de la conversation entière.

3. La fenêtre contextuelle — la table de travail

Analogie : la fenêtre contextuelle, c'est la table de travail de votre stagiaire IA. Tout ce qui n'est pas sur la table n'existe pas pour lui — peu importe ce que vous lui avez dit il y a 3 semaines ou 200 messages.

- **Une conversation trop longue dérive** : le début finit par sortir de la table. Symptôme : l'IA « oublie » vos consignes initiales.
- **Règle pratique** : une tâche = une conversation. Nouveau sujet → nouvelle conversation.
- **Coller 80 pages ≠ meilleure réponse** : l'IA lit moins bien le « milieu » des très longs documents. Donnez les extraits pertinents.
- **Pour reprendre un long travail** : demandez un résumé de la conversation, puis repartez à neuf avec ce résumé.

4. La mémoire — 3 niveaux distincts

NIVEAU	DURÉE DE VIE	VOUS LA CONTRÔLEZ ?
La conversation (fenêtre contextuelle)	Le temps d'un échange	Oui — coupez, résumez, redémarrez
La mémoire persistante (ChatGPT Memory, mémoire Claude)	Entre les sessions	Oui — consultez-la et faites le ménage dans les réglages
Instructions personnalisées et Projets	Permanente, choisie	Totalement — c'est la mémoire que VOUS rédigez

Réflexe pro : une fois par mois, ouvrez les réglages de mémoire de votre outil et supprimez ce qui est périmé ou que vous ne voulez pas qu'il retienne.

5. Choisir son modèle

VOTRE TÂCHE	TYPE DE MODÈLE	POURQUOI
Courriel, résumé, reformulation, brouillon	Modèle rapide	Instantané, largement suffisant
Analyse complexe, plan stratégique, problème à étapes	Modèle de raisonnement	Il « réfléchit » avant de répondre — plus lent, plus rigoureux
Question factuelle récente, veille	Modèle + recherche web	Sinon il répond avec des connaissances figées

- **Copilot ≠ ChatGPT** : même moteur, mais contraintes et contexte différents (Microsoft 365 vs ouvert).
- Les classements bougent chaque mois. Vérifiez plutôt que mémoriser : lmarena.ai et artificialanalysis.ai.

6. Votre devoir avant la séance 2

- Ouvrir les réglages de mémoire de mon outil et faire le ménage (supprimer ce qui est périmé ou indésirable).
- Tester la même tâche réelle sur 2 modèles différents (ex. : modèle rapide vs raisonnement) et noter la différence.

