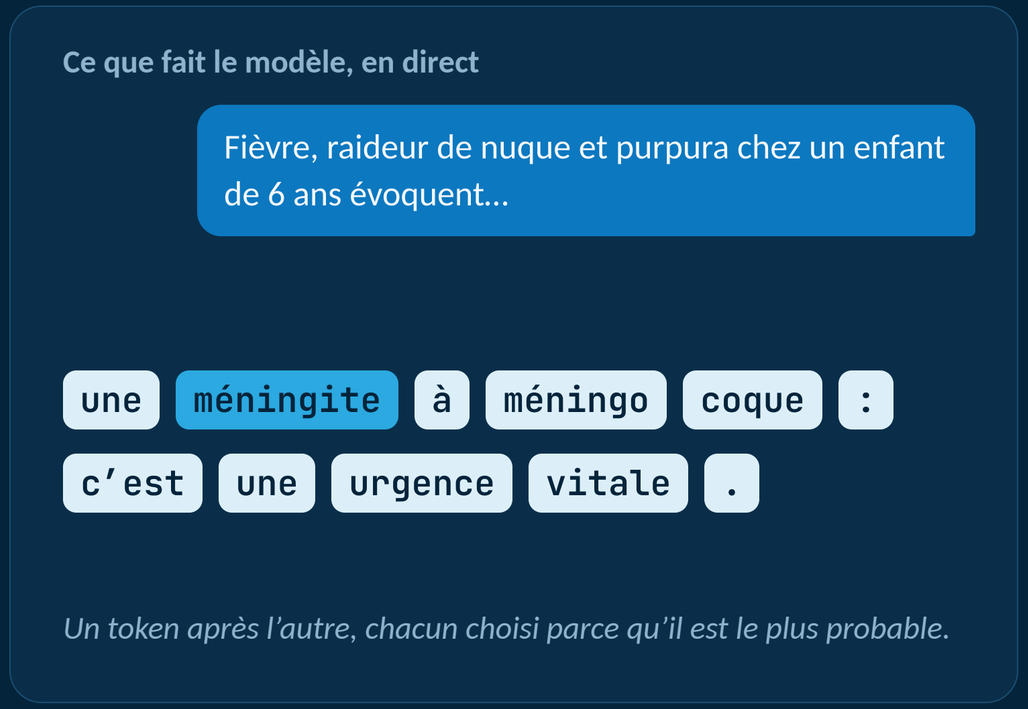
Formation IA et recherche scientifique · Soirée 1 · Modules 1 et 2

Comprendre les LLM

Comment une IA écrit, pourquoi elle se trompe, et comment bien lui parler

Une soirée pour les médecins, chercheurs et professionnels de santé : le fonctionnement des grands modèles de langage, leurs limites, puis le prompt engineering avancé sur des cas cliniques.



Formateur

Dr Boukary OUEDRAOGO

Médecin, PhD en systèmes d’information et épidémiologie · Fondateur de SYSTINFO

Lundi 9 novembre 2026

19h00 à 21h00 GMT · en direct sur Zoom

Formation intensive · 9 au 13 novembre 2026 · 19h à 21h GMT

Cinq soirées, de l’outil à la méthode

LUN 9 — **Comprendre les LLM et prompt engineering avancé** — Ce soir
Fonctionnement, limites, hallucinations · rôle, contexte, raisonnement guidé, gabarits

MAR 10 — **Recherche scientifique assistée**
Recherche bibliographique, Consensus, synthèse d’articles

MER 11 — **Revue systématique et méta-analyse**
PRISMA, extraction des données, risque de biais

JEU 12 — **Rédaction et intégrité scientifique**
IMRaD, citations, déclaration de l’usage de l’IA

VEN 13 — **Claude au poste de travail**
Tâches déléguées sur vos fichiers, Claude dans Excel, Word et PowerPoint

Les objectifs de ce soir

Quatre questions, une méthode

1

Comment un LLM fabrique-t-il une réponse ?

Tokens, prédiction, entraînement, mémoire de travail

2

Pourquoi invente-t-il parfois ?

Hallucinations, biais, limites en médecine

3

Quel modèle choisir, pour quel usage ?

Fermés, ouverts, outils de recherche

4

Comment lui poser la bonne question ?

Rôle, contexte, exemples, format, raisonnement guidé, gabarits

Puis à vous : douze exercices guidés sur **sante.systinfo.org**, avec l'IA de votre choix.

Où se trouve l'IA quand vous lui parlez ?

Répondez dans le chat : dans le cloud ? dans une base de données ? sur Internet ?

16
secondes pour répondre dans le chat

EXEMPLES DE RÉPONSES

R — Rasmata, sage-femme
Sur Internet, il cherche en direct sur PubMed

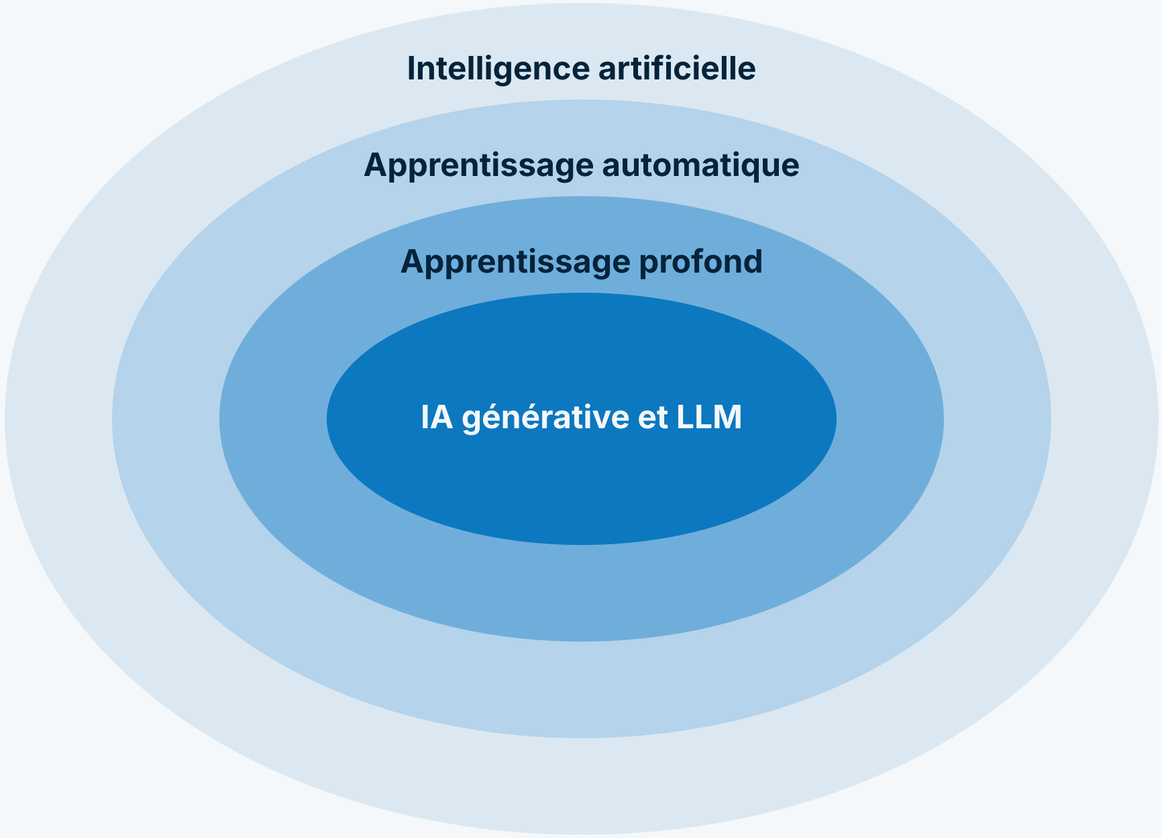
O — Pr Ousmane
Dans un supercalculateur

F — Fatim, doctorante
Dans mon téléphone ?

S — Dr Salif
Quelque part dans un data center

Indice : la réponse tient dans un fichier

L’IA en santé : des boîtes qui s’emboîtent



Intelligence artificielle
tout programme qui imite une tâche humaine
ex. : un système expert qui applique des règles de triage

Apprentissage automatique
les règles sont apprises à partir d’exemples
ex. : prédire le risque de réadmission à partir de dossiers

Apprentissage profond
des réseaux de neurones à plusieurs couches
ex. : lire une radiographie du thorax ou un frottis sanguin

IA générative et LLM
produit du texte, des images, du son
ex. : ChatGPT, Claude, Gemini rédigent un compte rendu

Large Language Model, grand modèle de langage

Un LLM, c’est quoi ?

Large

Des milliards de paramètres, réglés en lisant des milliers de milliards de mots : articles, livres, manuels, sites web.

Language

Il travaille sur du texte. Les modèles récents lisent aussi les images, les tableaux et la voix.

Model

Un modèle statistique : il calcule la suite la plus probable d’un texte. Il ne consulte pas une base de faits.

Imaginez un interne brillant qui aurait lu toute la bibliothèque médicale, sans avoir jamais examiné un patient, et sans savoir ce qu’il ne sait pas.

Très cultivé, très rapide, jamais responsable : c’est la bonne façon de le considérer.

Un modèle, c’est un fichier de nombres

LE MODÈLE DU PÉDIATRE

```
poids ≈ 2 × âge + 8
```

Estimation du poids d’un enfant de 1 à 10 ans en urgence

2

paramètres : 2 et 8

UN PETIT LLM

7 000 000 000

paramètres, soit 7 milliards de nombres rangés dans un fichier de quelques gigaoctets

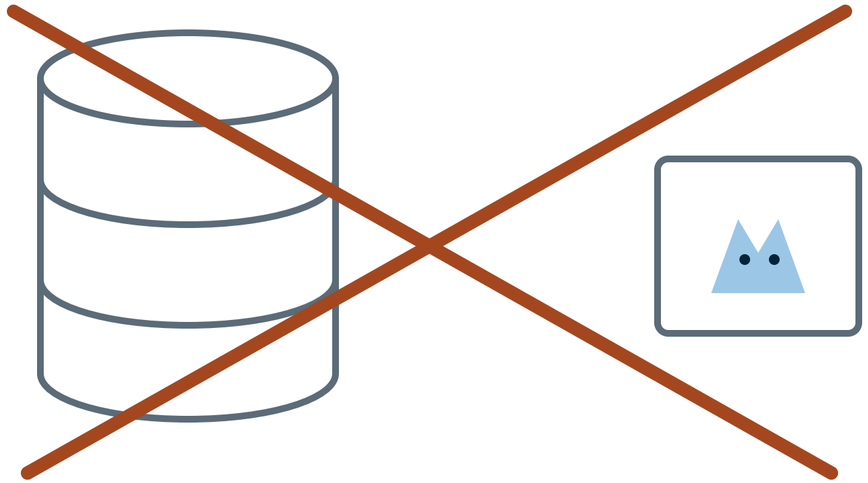
Le modèle ne stocke pas des réponses. Il stocke des régularités.

Il ne range rien : il reconstruit

~~pas de textes~~ ~~pas d'images~~ ~~pas de manuels~~ ~~pas de dossiers~~ **uniquement des régularités statistiques**

CE QU'IL NE FAIT PAS

aller chercher une photo de chat rangée quelque part



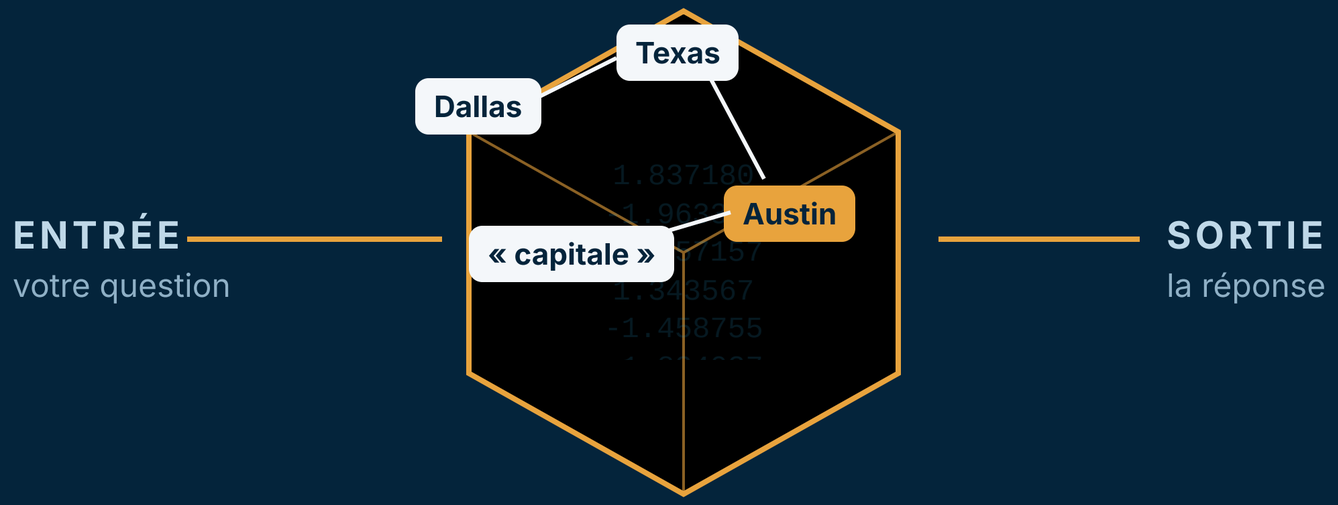
CE QU'IL FAIT

reconstruire un chat à partir de ce qu'il a compris des chats



Les nombres encodent ce que le modèle a constaté pendant sa fabrication. Pour un texte médical, c'est pareil : il reconstruit une réponse.

On sait le fabriquer, pas encore le lire



Impénétrable… pour l’instant

- ✓ **Le fabriquer**
- ✓ **L’utiliser**
- ? **Comprendre l’intérieur**

Les chercheurs commencent à peine

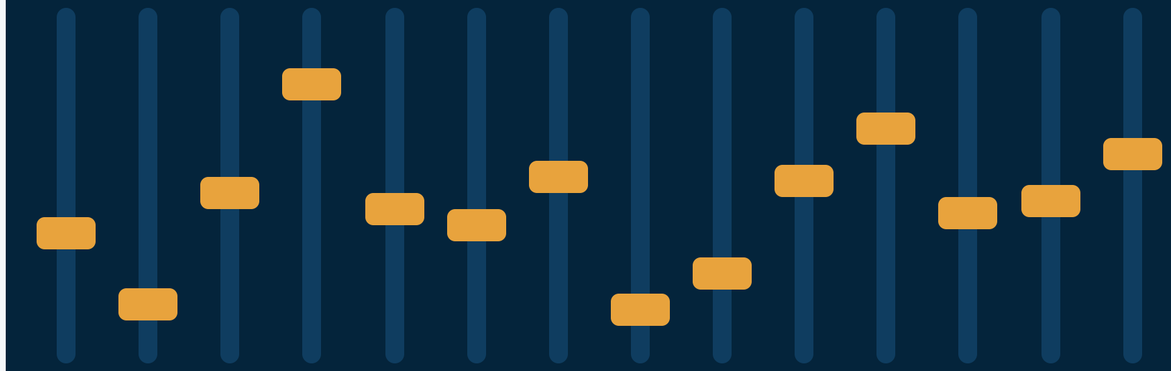
Anthropic, « On the Biology of a Large Language Model », mars 2025 : premiers « circuits » tracés dans Claude 3.5 Haiku, comme Dallas, puis Texas, puis capitale, puis Austin.

Comme le cerveau humain : on observe les entrées et les sorties, mais l’intérieur reste largement inconnu.

Apprendre, c’est ajuster des milliards de boutons

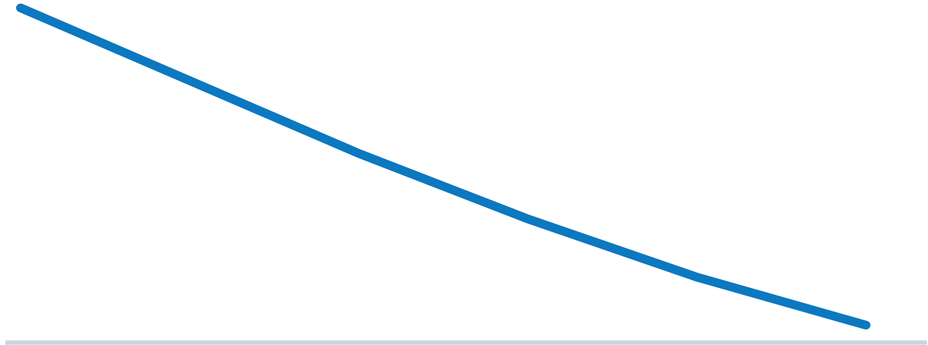
« Le paludisme est transmis par la piqûre d’un **moustique** » juste

LA TABLE DE MIXAGE : LES PARAMÈTRES



1 000 milliards de tours

L’ERREUR DIMINUE



erreur à chaque tour d’entraînement

Entraîner : des semaines sur des milliers de processeurs graphiques. Utiliser le modèle ensuite coûte infiniment moins.

Trois étapes, comme une formation médicale

Comment on fabrique un LLM

1

Pré-entraînement

Comme les études de médecine

Il lit des milliers de milliards de tokens et apprend à prédire la suite. Il acquiert la langue, des faits, des façons de raisonner. Des mois de calcul sur des milliers de processeurs.

2

Ajustement aux consignes

Comme l'internat

On lui montre des milliers d'exemples de consignes et de bonnes réponses : il apprend à répondre à une question plutôt qu'à prolonger un texte.

3

Alignement par retours humains

Comme la supervision par les seniors

Des évaluateurs comparent ses réponses : il apprend à préférer celles qui sont utiles, honnêtes et prudentes.

Résultat : un modèle figé à une date de coupure. Il ignore ce qui est paru après, sauf s'il peut chercher sur le web, et il ne voit jamais vos dossiers, sauf si vous les lui donnez.

Le modèle lit des jetons (tokens), pas des mots

CE QUE VOUS VOYEZ

hypercholestérolémie

20 lettres

CE QUE L'IA VOIT



5 jetons

Facturé au jeton par les services d'IA en ligne

+10 à 20 % de jetons en français qu'en anglais

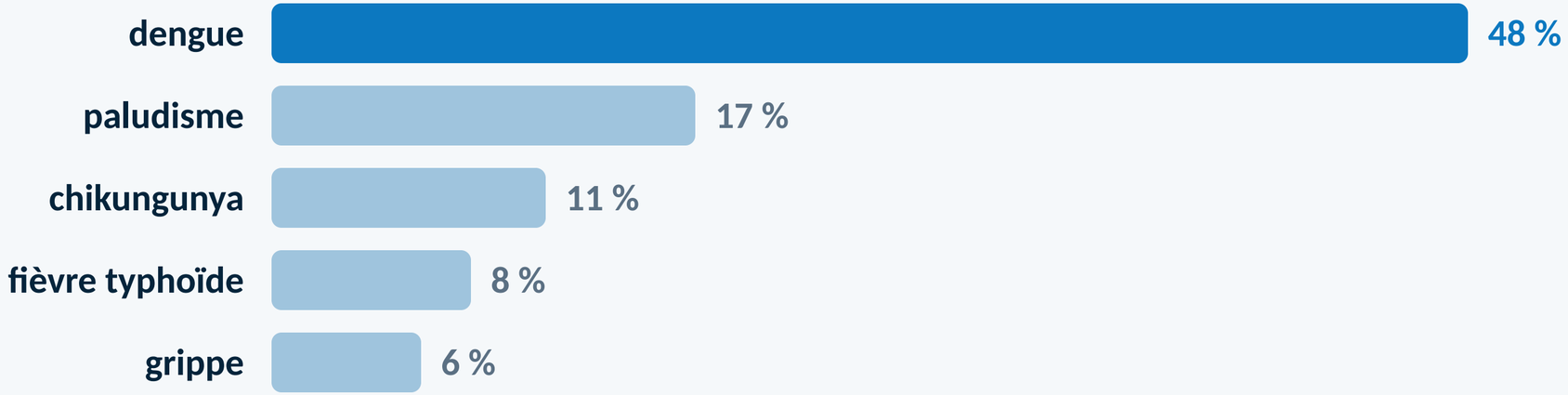
Bien plus en mooré, dioula ou fulfuldé

Découpages et numéros illustratifs : ils varient d'un modèle à l'autre.

Comment ça marche · étape 2

Il prédit le token suivant

```
Femme de 25 ans, Ouagadougou, fièvre à 39,5 °C depuis 3 jours, céphalées, douleurs articulaires. TDR paludisme négatif. Diagnostic le plus probable :
```



Il prend souvent le plus probable…

… mais pas toujours : une part de hasard (la température) lui fait parfois choisir le deuxième ou le troisième.

Probabilités illustratives. Le modèle ne « sait » pas : il calcule ce qui suit le plus vraisemblablement ce texte, d'après tout ce qu'il a lu.

Tout le secret : deviner le jeton suivant

Chez un enfant fébrile en zone d’endémie, le premier examen à réaliser est un TDR du paludisme.

Réponse terminée : **5 tours** de la même boucle

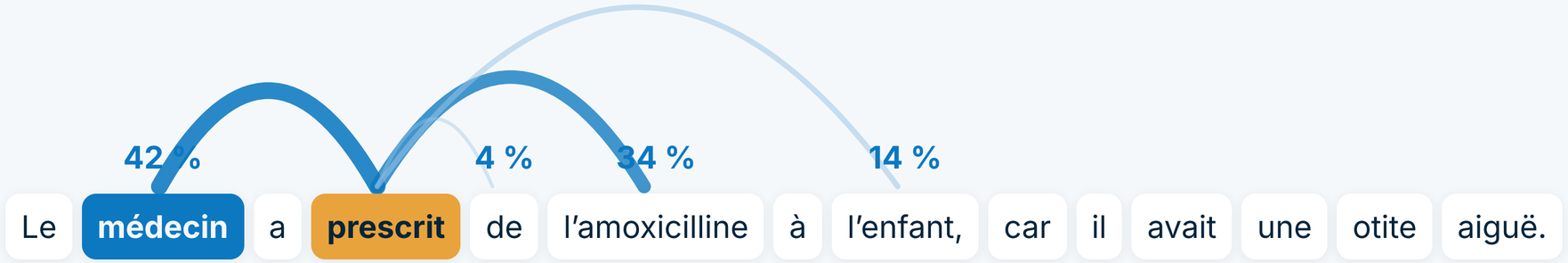
Probabilités illustratives



Température basse : il prend presque toujours le premier. **Haute** : il ose les suivants, plus créatif et plus risqué.

L’attention : chaque mot regarde les autres

Le mot **« prescrit »** cherche qui prescrit, quoi et à qui



Mécanisme inventé en 2017 dans l’architecture Transformer, le T de GPT. Pourcentages illustratifs.

Ce qu’on peut lui confier

Ce qu’il fait bien, ce qu’il ne peut pas savoir

Il fait bien

- Résumer un article ou une recommandation
- Reformuler pour un patient, traduire
- Structurer un compte rendu, un protocole
- Expliquer une notion classique
- Proposer des pistes, des plans, des questions

Il ne peut pas savoir

- Ce qui est paru après sa date de coupure
- Votre contexte : protocoles nationaux, médicaments disponibles
- Le dossier de votre patient, s’il ne l’a pas sous les yeux
- Ses propres erreurs : il ne sait pas ce qu’il ne sait pas
- La conversation d’hier, sauf mémoire activée

La fenêtre de contexte : sa mémoire de travail

Tout ce que le modèle « voit » au moment de répondre :



Grande, mais pas infinie

Selon les modèles, de quelques centaines de milliers à plus d’un million de tokens : d’un long rapport à plusieurs livres.

Le début pèse moins

Dans une très longue conversation, les consignes du début s’estompent. Nouvelle tâche, nouvelle conversation.

Tout part chez le fournisseur

Ce qui est dans la fenêtre est envoyé aux serveurs du modèle : on anonymise avant de coller.

Un réglage qui change la réponse

La température : constance ou créativité

Même question posée trois fois : « Propose un titre pour un article sur la dengue à Ouagadougou. »

Température basse (0,2)

- Dengue à Ouagadougou : épidémiologie et prise en charge
- Dengue à Ouagadougou : épidémiologie et prise en charge
- Dengue à Ouagadougou : épidémiologie et prise en charge

Température haute (1,0)

- Quand la fièvre n’est pas le paludisme : la dengue à Ouagadougou
- Dengue urbaine au Burkina Faso : les leçons d’une épidémie
- Moustiques, villes et fièvre : la dengue à Ouagadougou

Basse pour résumer, extraire, calculer. Plus haute pour imaginer des titres, des hypothèses, des messages.

Dans les applications grand public, on ne la règle pas : on la remplace par la consigne (« réponds de façon concise et constante »).

La nouvelle génération

Les modèles qui réfléchissent avant de répondre

Modèle rapide



Modèle de raisonnement



de 10 secondes à plusieurs minutes

Utile pour un diagnostic différentiel pédagogique, un calcul, une analyse statistique. Plus lent et plus cher.

Et pas infaillible : un raisonnement bien présenté peut être faux.

LA RÈGLE D'OR

L'IA produit le plausible, pas forcément le vrai

Et elle ne revient jamais en arrière : chaque jeton écrit devient une contrainte pour la suite.



Donner les sources

Fournir les recommandations, l'article, le protocole plutôt que compter sur sa mémoire



Vérifier

Exiger des références et contrôler chacune d'elles sur PubMed



Décider en clinicien

La responsabilité médicale ne se délègue pas à un outil

Limite n° 1

L’hallucination : faux, mais très convaincant

Référence proposée par une IA (exemple fictif)

Ouédraogo A, Sawadogo M, Kaboré J, et al. Efficacy of artesunate-amodiaquine in Burkinabe children: a randomised trial. *Lancet Infect Dis.* 2021;21(4):512-9.

INTROUVABLE SUR PUBMED

115 références produites par ChatGPT-3.5 dans des textes médicaux (Bhattacharyya et al., Cureus, 2023)

47 % inventées de toutes pièces

46 % réelles, mais inexactes (auteurs, année, pages…)

7 % exactes

Les modèles récents se trompent moins, mais le risque demeure : une référence non vérifiée n’entre jamais dans un document.

Comprendre pour s’en protéger

Pourquoi un LLM invente

Il vise le plausible, pas le vrai

Sa seule tâche : produire la suite la plus vraisemblable. Une fausse référence bien formée est très vraisemblable.

Il comble les trous

Maladie rare, donnée locale, publication récente : là où ses données sont pauvres, il extrapole.

La question oriente la réponse

« Quelle dose chez l’enfant ? » appelle un chiffre, même quand aucune dose pédiatrique n’est validée.

Il cherche à vous satisfaire

Si votre question suggère une réponse, il a tendance à la confirmer : c’est la complaisance.

Une hallucination a exactement l’allure d’une bonne réponse : même ton assuré, même mise en forme soignée.

Au-delà des hallucinations

Six limites à connaître en médecine

Biais des données

Corpus surtout anglophones et occidentaux : populations africaines, peaux foncées et maladies tropicales y sont sous-représentées.

Calculs et unités

Doses, conversions, clairances : il peut se tromper avec aplomb. On recalcule toujours.

Actualité

Les recommandations parues après sa date de coupure lui sont inconnues, sauf recherche web.

Confidentialité

Ce que vous collez part sur les serveurs du fournisseur. Jamais de données identifiantes.

Complaisance

Il a tendance à confirmer l'hypothèse que suggère votre question.

Responsabilité

Un assistant conversationnel n'est pas un dispositif médical validé : la décision reste au médecin.

Ce qu’on fait, concrètement

Réduire le risque : six réflexes

1. **Donner la source** : collez la recommandation ou l’article, et demandez de s’y limiter
2. **Autoriser le doute** : « si tu ne sais pas, dis-le », « signale ce qui manque »
3. **Vérifier chaque référence** : PubMed, DOI, site de la revue
4. **Préférer un outil branché sur les publications** pour la bibliographie : SIRA Recherche, Consensus, PubMed
5. **Croiser deux modèles** sur les questions importantes
6. **Garder la décision humaine**, et la trace : prompt, modèle, date

Novembre 2026

Le paysage des modèles

Fermés, en ligne

Les plus performants en général. Vos données partent chez le fournisseur.

- **ChatGPT · OpenAI**
- **Claude · Anthropic**
- **Gemini · Google**
- **Copilot · Microsoft**

Ouverts, installables

Téléchargeables : ils peuvent tourner sur un serveur local, même sans internet.

- **Llama · Meta**
- **Mistral · France**
- **DeepSeek · Chine**
- **Qwen · Alibaba**

Outils de recherche

Branchés sur les publications : chaque affirmation renvoie à un article.

- **Consensus**
- **Elicit**
- **OpenEvidence**
- **SIRA Recherche · SYSTINFO**

Chaque nom recouvre plusieurs versions, du plus rapide au plus puissant, renouvelées tous les quelques mois : on juge sur des tests, pas sur la marque.

Choisir son modèle

Quel outil pour quel usage ?

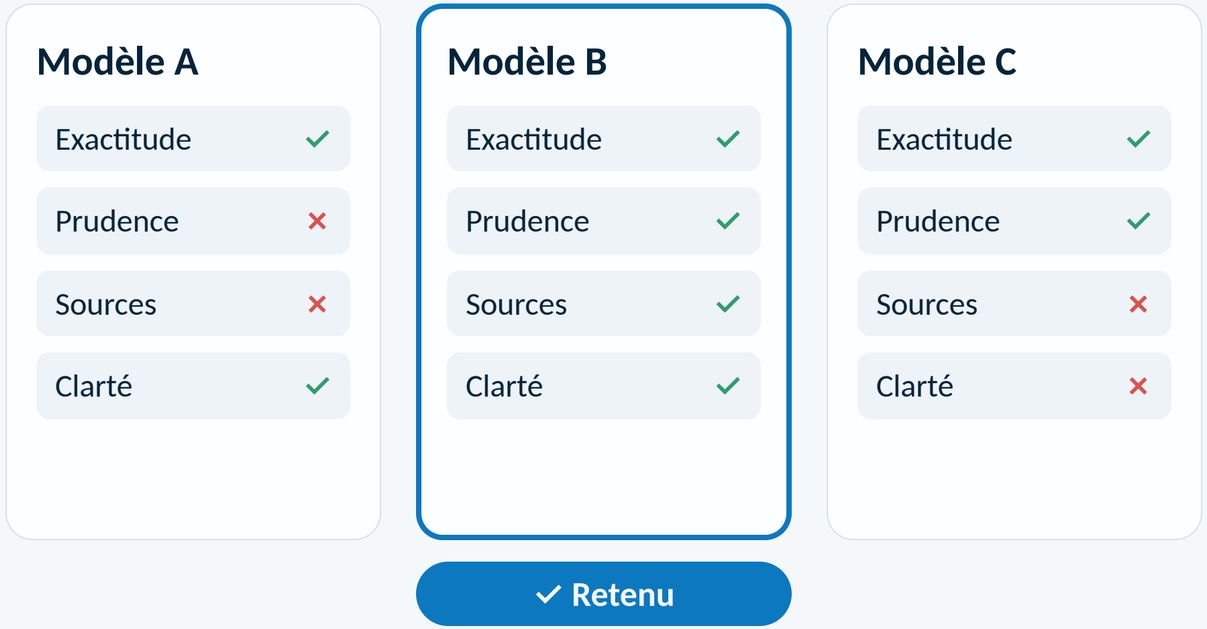
Votre besoin	Type d'outil	Pourquoi
Résumer un article, reformuler, traduire	Modèle rapide	Tâche simple, coût faible
Diagnostic différentiel pédagogique, calcul, statistiques	Modèle de raisonnement	Étapes visibles, donc vérifiables
Recherche bibliographique	Outil branché sur les publications	Références réelles, avec leurs liens
Données sensibles : dossiers, cohortes	Modèle local ou hébergement maîtrisé	Les données ne sortent pas
Langues nationales	Outil avec traduction dédiée	Les grands modèles y restent faibles

Critères à peser : confidentialité, qualité du raisonnement, accès aux sources, langue, coût, disponibilité dans votre pays.

Une méthode simple

Comparer avant d’adopter

1. Prenez trois cas fictifs de votre pratique
2. Posez-les à deux ou trois modèles, sans regarder les noms
3. Notez : exactitude, prudence, sources, clarté
4. Gardez le meilleur pour cet usage précis



Dans SIRA, le bouton ⚖️ Comparer pose le même prompt à plusieurs modèles côte à côte : 50 crédits offerts sur sira.systinfo.org.

Une question vague donne une réponse vague

Question vague

```
Parle-moi de la dengue.
```

Ce que vous obtenez

Un cours général sur la maladie, sans lien avec votre patient.

Question précise

```
Je suis médecin généraliste dans un CSPS de Ouagadougou.
Patiente fictive de 25 ans : fièvre à 39,5 °C depuis 3 jours,
céphalées, douleurs articulaires, TDR paludisme négatif. Liste
les 3 diagnostics les plus probables, les signes d'alerte à
rechercher et les examens disponibles au premier niveau. Réponds
en 10 lignes.
```

Ce que vous obtenez

Une réponse ciblée, utilisable en consultation, et vérifiable.

Ce qui change : le contexte, la tâche, le format. Rien de technique.

Un prompt avancé, brique par brique

Rôle	`Tu es pharmacien clinicien dans un CHU, habitué à relire les ordonnances des patients âgés.`
Contexte	`Patient fictif de 78 ans, insuffisance rénale modérée, fibrillation atriale. L'ordonnance est collée ci-dessous.`
Tâche	`Repère les interactions et les adaptations de dose à vérifier.`
Contraintes	`Appuie-toi uniquement sur l'ordonnance fournie. Si une information manque, dis-le.`
Format	`Un tableau : médicament, problème, gravité, action proposée. Puis 3 lignes de synthèse.`
Vérification	`Relis ton tableau et signale tout point dont tu n'es pas certain.`

Six questions qu'un confrère vous poserait avant de vous aider : qui je suis, sur quoi, pour faire quoi, avec quelles limites, sous quelle forme, avec quel contrôle.

Brique 1 · le rôle

Le rôle oriente le registre, pas le savoir

Rôle vague

```
Tu es un expert en médecine.
```

Un ton assuré, un contenu général : le rôle n’apporte presque rien.

Rôle précis

```
Tu es pharmacien clinicien en
CHU. Tu relis les ordonnances
des patients âgés en
insuffisance rénale.
```

Les priorités, le vocabulaire et la vigilance d’un métier précis.

Rôle inversé

```
Avant de répondre, pose-moi une
question à la fois jusqu’à
avoir ce qu’il te faut.
```

L’IA vous interroge comme un confrère avant de se prononcer.

Le rôle ne donne au modèle aucune connaissance nouvelle : il choisit, dans ce qu’il sait déjà, le ton, les priorités et le niveau de détail.

Brique 2 · le contexte

Donner la source, puis l’y enfermer

```
Tu réponds uniquement à partir du texte placé entre les balises
<protocole>.
Si la réponse n’y figure pas, écris : « Non précisé dans le protocole ».
Cite le passage exact qui justifie chaque réponse.

<protocole>
[coller ici le protocole national de prise en charge]
</protocole>

Question : quelle est la conduite à tenir devant un paludisme grave chez
l’enfant de moins de 5 ans ?
```

1. **Délimiter**
   Des balises ou des triples guillemets séparent la source de la consigne.

2. **Restreindre**
   « Uniquement à partir de ce texte » : le modèle ne puise plus dans sa mémoire.

3. **Faire citer**
   Chaque réponse renvoie au passage exact : vous vérifiez en dix secondes.

Un modèle qui travaille sur une source fournie invente beaucoup moins. C’est le premier réflexe de la règle d’or.

Brique 3 · les exemples

Deux exemples valent mieux qu’une consigne

```
Classe chaque motif d’appel en : URGENCE, CONSULTATION DU JOUR ou
CONSEIL.

Motif : « Enfant de 2 ans, fièvre et convulsions depuis 10 minutes »
Classe : URGENCE

Motif : « Toux sèche depuis 2 jours, sans fièvre, adulte en bonne santé »
Classe : CONSEIL

Motif : « Femme enceinte de 32 semaines, céphalées intenses et vision
floue »
Classe :
```

Sans exemple

Le modèle devine les catégories, le format et le niveau d’exigence.

Avec deux ou trois exemples

Il imite le format, les catégories et le seuil de gravité que vous avez montrés.

Attention

Il imite aussi vos erreurs : choisissez des exemples justes et variés.

Réponse attendue : URGENCE

Signes de gravité possibles d’une pré-éclampsie : l’exemple montre le seuil attendu.

Brique 4 · le format

La sortie se commande

Tableau

« Colonnes : auteur, année, pays, devis, effectif, résultat principal. »

Longueur

« 150 mots au plus », « 5 puces », « une phrase par idée ».

Structure

« Introduction, méthodes, résultats, discussion », « SOAP » pour une observation.

Format machine

« Réponds en JSON avec les clés age, sexe, diagnostic, gravite. »

Un format fixé rend la réponse comparable d'un essai à l'autre, d'un modèle à l'autre, et vérifiable d'un coup d'œil.

C'est aussi ce qui permet de coller la sortie directement dans Excel, un tableau d'extraction ou un rapport.

Quatre techniques, quatre usages cliniques

CO-STAR

Contexte, objectif, style, ton, audience, réponse

Rédiger pour un public précis

Fiche patient, lettre, compte rendu

CoT

Chain of Thought, chaîne de pensée

Faire raisonner étape par étape

Calcul de dose, clairance, statistiques

ToT

Tree of Thoughts, arbre de pensées

Explorer plusieurs pistes, garder la meilleure

Diagnostic différentiel, options thérapeutiques

Chaînage

Prompt chaining, prompts en chaîne

Découper une tâche longue en étapes

Revue de littérature, protocole, article

À combiner avec les briques : rôle, contexte, exemples, format et vérification.

Technique 1 · CO-STAR

Six rubriques pour une fiche patient

C **Contexte**

« Médecin généraliste, consultation de suivi, patient de 60 ans, hypertension récemment découverte »

O **Objectif**

« Rédige une fiche d'information à lui remettre »

S **Style**

« Phrases courtes, mots simples, sans jargon »

T **Ton**

« Rassurant et encourageant »

A **Audience**

« Patient peu scolarisé, fiche lue à voix haute par un proche »

R **Réponse attendue**

« 150 mots, puis 5 conseils pratiques numérotés »

Pourquoi ça marche : chaque rubrique répond à une question que le modèle ne peut pas deviner. Le prompt complet est dans l'exercice 6.

Technique 2 · CoT, Chain of Thought

Faire raisonner étape par étape

```
Patient fictif de 70 ans, 60 kg, créatinine 150 µmol/L. Calcule la clairance de la créatinine selon Cockcroft-Gault, étape par étape, puis dis si une adaptation de dose est à vérifier. Montre chaque calcul.
```

1. Formule (homme) : (140 − âge) × poids × 1,23 / créatinine en µmol/L
2. (140 − 70) × 60 × 1,23 = 5 166
3. 5 166 / 150 ≈ **34 mL/min**
4. Clairance inférieure à 50 mL/min : vérifier l'adaptation de dose dans le RCP du médicament

Avec les étapes, une erreur se voit et se corrige. Sans elles, on reçoit un chiffre à croire sur parole.

Technique 3 · ToT, Tree of Thoughts

Trois experts, une décision

Patiente de 25 ans : fièvre, céphalées, arthralgies, TDR paludisme négatif

Expert 1 · Dengue

Pour : arthralgies, TDR négatif, contexte urbain

Contre : à confirmer (NS1)

Expert 2 · Paludisme malgré le TDR

Pour : zone d'endémie, faux négatifs possibles

Contre : TDR négatif

Expert 3 · Infection bactérienne

Pour : fièvre élevée

Contre : aucun foyer évident

Retenu : dengue, à confirmer par NS1 · En réserve : paludisme (goutte épaisse) · À surveiller : les signes d'alerte

La valeur du ToT : la piste en réserve et l'examen qui permet de trancher. La décision reste clinique et biologique.

Raisonnement guidé · au-delà de « étape par étape »

Brouillon, critique, version finale

1

Brouillon

« Rédige une première version de la fiche. »

2

Critique

« Relis-la comme un relecteur exigeant : erreurs, oublis, affirmations sans source. Liste-les. »

3

Version finale

« Corrige en tenant compte de ta critique et indique ce qui a changé. »

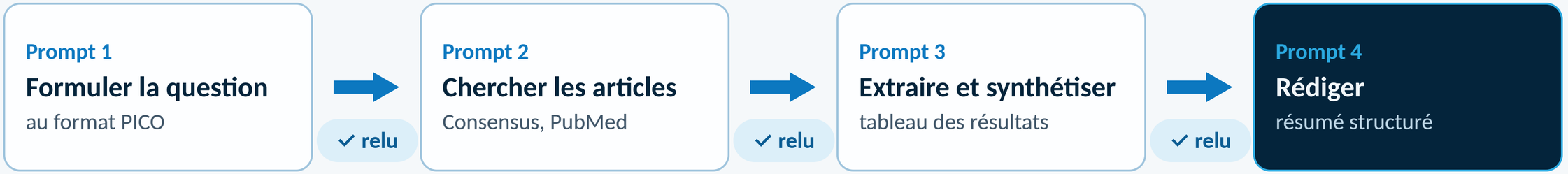
Variante à deux rôles : le rédacteur propose, le relecteur attaque (pharmacien, méthodologiste, comité d'éthique).

Ou bien : « Avant de conclure, liste les hypothèses que tu fais et les données qui te manquent. »

L'autocritique attrape des erreurs et améliore la forme, mais elle ne remplace pas une vérification externe : PubMed, RCP, recommandations.

Technique 4 · chaînage de prompts

Une tâche longue, plusieurs maillons



La réponse de chaque prompt devient l'entrée du suivant, après relecture.

Une étape, un prompt

Chaque maillon a un seul objectif : plus simple à écrire, plus simple à vérifier.

Je relis entre deux maillons

Une erreur corrigée tôt ne se propage pas jusqu'au texte final.

Le fil de la semaine

Mercredi la recherche, jeudi la revue systématique, vendredi la rédaction.

Gabarits réutilisables

Écrire une fois, réutiliser chaque semaine

```
GABARIT · Résumé d’article pour le staff

Rôle : tu es [spécialité] et tu prépares le staff du [jour].
Source : l’article placé entre les balises <article>.
Tâche : résume-le pour [public] en [nombre] mots.
Format : question, méthode, résultat principal chiffré,
limites, ce que cela change pour notre pratique.
Règle : n’utilise que l’article ; écris « non précisé » si
besoin.

<article>[coller le résumé ou le texte intégral]</article>
```

Où ranger vos gabarits

- ✓ Un document Word partagé par le service : la bibliothèque d’équipe
- ✓ Un Projet dans Claude ou ChatGPT : instructions et documents de référence gardés
- ✓ Un assistant personnalisé : GPT, Gem de Gemini
- ✓ SIRA, pour tester le même gabarit sur trois modèles avec ⚖️ Comparer

Les crochets sont les seules parties à changer d’une fois à l’autre.

Le raccourci

Faire écrire le prompt par l’IA

```
Je veux un prompt pour [préparer une fiche
d’éducation thérapeutique sur le diabète de type 2
pour des patients peu scolarisés].

Pose-moi d’abord, une à une, les questions nécessaires :
public, contexte, format, contraintes.

Puis rédige le prompt complet avec ces rubriques :
rôle, contexte, tâche, contraintes, format, vérification.
```

Comment faire

1. Décrivez le besoin en une ou deux phrases
2. Répondez aux questions de l’IA
3. Relisez le prompt proposé, corrigez, testez
4. Gardez-le comme gabarit s’il marche

Dans SIRA : domaine Prompt engineering, puis les boutons CO-STAR, COT et TOT.

Vous gardez la main : c’est vous qui savez ce que la tâche exige.

Garde-fous

L’IA prépare, le médecin décide

Avant d’envoyer

- Anonymiser : ni nom, ni date de naissance, ni numéro de dossier, ni photo identifiante
- Ne donner que les données utiles à la question
- Savoir où sont hébergées les données : outil institutionnel, modèle local

Avant d’utiliser la réponse

- Vérifier doses, interactions et références
- Confronter aux recommandations nationales
- Garder la trace : prompt, modèle, date
- Déclarer l’usage de l’IA dans une publication (vendredi)

L’IA accélère la préparation. La responsabilité reste celle du médecin.

À vous de jouer

Douze exercices, ce soir et toute la semaine

1 à 3	**Voir le modèle fonctionner**	Prédiction, tokens, température
4 et 5	**Débusquer une hallucination**	Références à vérifier, étude inventée
6 à 9	**Premiers prompts structurés**	CO-STAR, CoT, ToT, chaînage, sur des cas fictifs
10 à 12	**Prompt engineering avancé**	Rôle inversé, source et exemples, gabarit et autocritique

sante.systinfo.org

Les exercices, la présentation animée et le PDF

Avec l'IA de votre choix (ChatGPT, Claude, Gemini, Mistral…) ou avec SIRA, 50 crédits offerts sur sira.systinfo.org. Cas fictifs uniquement.

Merci pour votre attention

Place aux questions

Demain, mardi 10 novembre, 19h GMT : la recherche scientifique assistée.

Dr Boukary OUEDRAOGO

boukary@systinfo.ai · support@systinfo.ai · +226 25 45 46 43

Trois choses à retenir

1. Un LLM prédit le texte le plus probable : il ne sait pas, il calcule.
2. Une hallucination ressemble à une bonne réponse : on vérifie tout.
3. Un bon prompt donne un rôle, une source, un format et un contrôle, puis se garde comme gabarit.