

Ce qu'elle sait faire, et ce qu'elle ne peut pas savoir

Un seul mécanisme à comprendre, et tu sais quoi lui confier

Module F1 - COMPRENDRE - 11 min - version 3.1

Une IA générative - ChatGPT, Claude, Gemini - écrit avec le même aplomb qu'elle reformule ton texte ou qu'elle invente une date. Ce module te donne la ligne qui sépare les deux, et le contrôle à faire de chaque côté.

Ce que tu sauras faire en sortant

- Dire en une phrase comment une IA générative fabrique sa réponse
- Trier une tâche en trois cas : transformation, production, ou les deux à la fois
- Estimer ce qu'une erreur coûterait, et en déduire ton niveau de contrôle
- Écrire une consigne qui l'empêche d'ajouter ce que tu ne lui as pas donné

Teste-toi avant de commencer

Reponds dans ta tête avant de lire la réponse. Si les trois étaient déjà les tiennes, tu sais l'essentiel de ce module et tu peux passer au suivant. Rien n'est noté.

- Tu joins deux contrats et tu lui demandes de les comparer, puis de te recommander le meilleur. Qu'est-ce qui, là-dedans, se vérifie en relisant, et qu'est-ce qui se vérifie en ouvrant une source ?
- Réponse : La comparaison transforme des documents que tu as fournis : elle se relit contre eux. La recommandation, elle, est un arbitrage qui reste signé par toi. Et une garantie citée qui ne figure dans aucun des deux contrats est une production : elle se retire, elle ne se vérifie pas.
- Même reformulation, deux destinations : un brouillon qui reste sur ton disque, un devis qui part chez un client. Ton contrôle change-t-il ?
- Réponse : Oui, et c'est la seule chose qui change. Même tâche, même risque de mot durci - mais un devis peut t'être opposé plus tard. Le test tient en une phrase : si quelqu'un pouvait me l'opposer - un client, un assureur, un contrôleur - je traite le point comme un impact élevé.
- Deux consignes possibles : « sois factuel, n'invente rien », ou « n'ajoute aucun chiffre, délai, nom ni référence que je ne t'ai pas donné ; écris à confirmer à la place ». Laquelle t'évite une invention ?
- Réponse : La seconde. « N'invente rien » ne nomme rien de contrôlable : le modèle n'a aucun moyen de savoir ce qu'il ignore. La seconde nomme les catégories qu'il n'a pas le droit de produire et lui donne une sortie quand l'information manque - c'est ce qui rend le trou visible au lieu de le combler.

Vérifie ce qui te reste (à faire après lecture)

- Ta note dit « le remplacement peut être nécessaire ». La reformulation revient avec « doit être effectué ». Rien n'a été ajouté : tu laisses passer ?
- Réponse : Non. Rien n'a été inventé, et pourtant une réserve est devenue une obligation - si ce texte part chez le client, c'est toi qui l'as écrite. Le signal à chercher n'est pas un élément en trop, c'est un mot qui a durci.
- A relire : Le piège : une transformation peut déformer
- Tu demandes un délai légal sans rien fournir, et la réponse tombe identique trois fois de suite. Cette

stabilité te dit-elle qu'elle est juste ?

- Reponse : Non. Une formulation figée est écrite partout de la même façon, donc elle ressort stable - y compris dans sa version d'avant la dernière réforme. La stabilité dit que la formulation était fréquente, rien de plus. Le seul contrôle : où est-ce publié, et est-ce que j'ouvre la page ?
 - A relire : Comment ça marche ?
 - Tu relis la réponse et tout est exact. Reste-t-il quelque chose à contrôler ?
 - Reponse : Oui : d'où vient chaque élément. « Exact » se juge contre une source, et tu ne l'as pas ouverte pour ce que tu n'as pas fourni. La question ne se pose pas sur la tâche, elle se pose ligne par ligne.
 - A relire : Première décision : qui fournit la matière ?
-

L'essentiel, en trente secondes

Une IA générative apprend des régularités dans d'énormes quantités de texte. Une tournure, une structure, un ton : c'est régulier, elle est excellente. Un montant, une date, le nom d'un fournisseur local : c'est arbitraire, il n'y a aucune régularité à apprendre. Elle produit quand même, et la forme reste impeccable.

La forme ne t'indique donc jamais de quel côté tu te trouves.

D'où la règle du module : dans les deux cas tu contrôles, mais pas la même chose.

- La matière vient de toi tu vérifies ce qu'elle en a fait.
- La matière vient d'elle tu vérifies d'où ça sort.

Puis tu ajustes l'effort selon ce qu'une erreur coûterait.

La situation

Tu ouvres ChatGPT et tu lui fais reformuler ton devis en langage clair : c'est bon du premier coup, et tu n'y as pas passé ta soirée.

Dans la foulée, tu lui demandes le délai pour contester une facture. La réponse arrive, aussi bien tournée que la première.

Même outil, même aplomb, deux fiabilités qui n'ont rien à voir.

Comment ça marche ?

Complète à voix haute : « Je vous prie d'agréer, Madame, Monsieur, l'expression de... »

Tu as fini sans réfléchir, sans consulter aucune base de politesses : tu as vu cette formule cent fois. C'est une image, pas une description de la machine - mais ce qu'elle a en commun avec ta phrase suffit ici : un travail sur des régularités apprises, pas une recherche dans un registre.

Deux conséquences, et elles portent tout le parcours.

Ce qu'elle « sait » est rangé dans le modèle, pas dans une base qu'on interroge. Il n'y a pas de fiche à ouvrir, pas de source à citer.

Certains outils vont chercher des documents avant de répondre. Ça change ce qu'ils ont sous les yeux, pas leur façon d'écrire la suite. Une source réduit le risque, elle ne le supprime pas.

Et une réponse qui tombe pareil à chaque essai n'est pas pour autant vérifiée. Un délai de garantie, un seuil réglementaire, un tarif conventionnel : ce sont des formulations figées, écrites partout de la même façon. Elles ressortent stables - y compris dans leur version d'avant la dernière réforme. La stabilité dit que la formulation était fréquente, jamais qu'elle est en vigueur.

Première décision : qui fournit la matière ?

En une phrase. Une même réponse contient presque toujours les deux : ce que tu as fourni et ce qu'elle a comblé. Tu relis la première partie en cherchant un mot qui a durci, et tu vérifies la seconde en ouvrant la source.

Trois mots, et le module ne parle plus que d'eux :

- Tu fournis ton propre texte et demandes de le raccourcir transformation.
- Tu demandes une date, une référence ou un nom que tu n'as pas fourni production.
- Tu fournis des notes et demandes aussi de compléter ce qui manque tâche mixte.

Le cas mixte est le plus fréquent, et c'est celui qu'on rate. Une tâche entière est rarement de l'un ou de l'autre. La question ne se pose donc pas sur la tâche, elle se pose ligne par ligne, sur l'origine de chaque élément.

Le piège : une transformation peut déformer

C'est le point qu'on retient le moins, et c'est le plus coûteux. Reformuler ne consiste pas à déplacer des mots : il faut choisir lesquels. Et un mot qui change peut changer un engagement.

Ta note de visite dit :

« Le remplacement du tableau électrique peut être nécessaire selon l'état des câbles. »

La reformulation revient :

« Le remplacement du tableau électrique doit être effectué. »

Rien n'a été inventé - le tableau était bien dans ta note. Mais une réserve est devenue une obligation, et si ce texte part chez le client, c'est toi qui l'as écrite.

Le signal à chercher n'est donc pas un élément en trop, c'est un mot qui a durci : peut devenu doit, environ disparu, à confirmer évaporé, un délai indicatif devenu ferme.

Relis en te demandant non pas « est-ce que c'est faux ? » mais « est-ce que je m'engage plus qu'avant ? »

Deuxième décision : que coûte l'erreur ?

| Impact de l'erreur | Transformation - la matière vient de toi | Production - l'information vient d'elle |
|---|---|---| | Faible - brouillon qui reste chez toi | Relecture rapide contre ta matière | Utiliser comme piste ; vérifier les faits seulement s'ils sont conservés | | Moyen - contenu diffusé ou envoyé | Comparer engagements, chiffres, dates et noms avec la matière fournie | Vérifier chiffres, dates, noms et références dans une source adaptée | | Élevé - décision ou engagement juridique, financier, sécurité | Comparaison ligne à ligne avec la source, et validation humaine compétente | Source officielle et contrôle humain compétent avant décision |

Cas mixte : applique la colonne « transformation » à ce que tu as fourni, et la colonne « production » à tout élément nouveau. Le contrôle final est le plus exigeant des deux.

Entre moyen et élevé, le test est simple : si quelqu'un pouvait me l'opposer plus tard - un client, un assureur, un contrôleur - traite le point comme un impact élevé.

Ce classement est une recommandation Solutio, pas une mesure de fréquence.

Reste l'arbitrage, qu'on ne lui délègue pas - non qu'elle en soit incapable, mais parce que la décision reste signée par toi.

L'exemple guidé

Tu lui demandes : « Rédige une réponse à ce client. Dis-lui que le devis reste valable 30 jours et que la livraison aura lieu sous deux semaines. »

Ce que tu as réellement fourni : le prix, les prestations, le nom du client. Ce qui n'y était pas : la validité de 30 jours, le délai de livraison.

Le tri. La structure de la réponse transforme ta matière. Les deux délais, eux, sont des productions : tu les as énoncés dans ta consigne, mais rien dans ton dossier ne les établit.

La décision. Ces deux délais doivent être confirmés depuis ta source ou retirés du message. Ils ne se relisent pas, ils se vérifient.

L'impact. Ça part chez un client et un devis peut t'être opposé plus tard : impact élevé. Donc source, et relecture des mots qui ont durci avant envoi.

Un prompt à emporter

PROPOSITION - NON TESTÉE / NON REJOUÉE AU BANC. Ce bloc applique le tri du module ; il ne constitue pas une preuve d'efficacité.

`` Utilise uniquement les informations que je fournis ci-dessous.

Avant de rédiger : 1. sépare les éléments explicitement fournis de ceux qui manquent ; 2. indique les informations nécessaires pour répondre correctement ; 3. n'ajoute aucun chiffre, délai, nom, règle, engagement ou référence absent ; 4. marque « à confirmer » lorsqu'un point ne peut pas être établi à partir de ma matière.

Ensuite, produis le document demandé. `

Le prompt ne remplace ni la source ni la relecture.

À toi

Prends trois tâches écrites de ta semaine. Pour chacune, deux lignes : matière fournie ou information demandée, puis ce que ça casse si c'est faux.

Avant de coller quoi que ce soit. Pour ce premier essai, choisis un texte que tu pourrais afficher en vitrine : pas de nom de personne, pas de coordonnées, pas de montant qui identifie quelqu'un, rien qui touche à la santé ou au dossier d'un salarié. Si ton texte en contient, remplace-les par des étiquettes - [CLIENT], [MONTANT] - à la main, avant d'ouvrir l'outil. Jamais en le lui demandant : pour qu'il masque, il faudrait d'abord tout lui envoyer. Ce tri est le sujet entier de « Ce que tu as le droit de mettre dedans » ; d'ici là, tiens-toi à cette règle.

Prends ensuite la plus clairement « matière fournie », et fais-la faire : ton premier essai est choisi pour réussir.

Tu as réussi si :

- tu peux dire, pour chaque information de la réponse, si elle vient de toi ou d'elle ;
- tu as relu en cherchant les mots qui ont durci - un peut devenu doit, un environ disparu - et pas seulement les éléments ajoutés ;
- tu as nommé ce que l'erreur coûterait ;
- tu sais quoi contrôler avant d'envoyer.

Si tes trois tâches tombent du même côté. Toutes en transformation : commence aujourd'hui, « Bien lui parler, pour arrêter d'obtenir du tiède » t'y aidera. Toutes en production : ce n'est pas le bon point d'entrée, cherche une tâche de mise en forme et garde tes recherches pour après « Repérer quand elle invente ».

Pas envie de fouiller tes archives ? Trie ces trois-là - à écrire sans aucune donnée réelle : relancer une facture impayée · résumer un compte-rendu de chantier · rédiger une annonce de recrutement.

Dans huit jours : reprends ta liste, ajoute la tâche que tu as réellement confiée, et regarde si ton tri tenait.

Selon ce que tu obtiens

Ta demande a produit un texte utilisable. Regarde surtout s'il contient un élément que tu n'avais pas donné - c'est là que se joue la suite du parcours, avec « Repérer quand elle invente ».

Le texte est correct mais tiède. Ce n'est presque jamais la formulation : c'est la matière. Ajoute le mail auquel tu réponds, ou tes notes, et relance.

Le texte invente une information sur ton activité. Bon signal, tôt : tu viens de voir en direct ce que « il prédit la suite » veut dire.

Tu n'as pas su quoi lui demander. Reprends la question la plus répétitive de ta semaine - celle que tu réécris à chaque fois. C'est toujours le bon premier essai.

Micro-défi

Trois demandes. Une seule est mal adressée à une IA générative. Laquelle ?

- « Réécris ce mail de refus pour qu'il soit plus court et moins sec. »
- « Combien me reste-t-il sur ma facture Dupont ? »
- « Donne-moi trois façons de présenter cette offre à un client pressé. »

Réponse attendue : la 2. Elle demande un fait qui vit dans ton outil de gestion, pas dans un texte - le modèle ne peut que le deviner. Les deux autres demandent de la mise en forme et des variantes, ce qu'il fait bien.

Fiche mémo

Ce qu'une IA générative sait faire, et ce qu'elle ne sait pas

- Elle prédit du texte plausible - elle ne consulte pas tes dossiers.
- Elle est bonne sur : reformuler, raccourcir, structurer, proposer des variantes.
- Elle est mauvaise sur : tes chiffres, tes dates, tes faits internes.
- Plus la demande contient de matière à toi, meilleur est le résultat.
- Ce qu'elle ajoute et que tu n'as pas fourni est à vérifier, toujours.

- Une réponse assurée n'est pas une réponse vérifiée.

Ce que tu viens de pratiquer

- Tu sais d'où vient une réponse : des régularités apprises, pas une base qu'on interroge
- Tu tries une tâche en deux questions : qui fournit la matière, et ce que l'erreur coûterait
- Tu sais qu'une transformation aussi peut déformer, et quel mot chercher en relisant

La suite : « Bien lui parler, pour arrêter d'obtenir du tiède » pour lui donner la matière, « Ce que tu as le droit de mettre dedans » pour ce qu'on a le droit d'y mettre, « Repérer quand elle invente » pour repérer ses inventions.

Pour aller plus loin

Pourquoi c'est vrai

Pendant son apprentissage, le modèle apprend « la distribution de la langue dans un grand corpus de textes » (Kalai et coll., OpenAI, 2025 (<https://arxiv.org/abs/2509.04664>), traduit de l'anglais). C'est ce qu'on appelle ici des régularités.

Que la forme ne dise rien de la justesse est ma déduction, pas un résultat mesuré. Ce qui est mesuré est plus étroit : en 2020, des évaluateurs humains distinguaient difficilement des articles de presse produits par GPT-3 de ceux écrits par des humains (Brown et coll. (<https://arxiv.org/abs/2005.14165>), traduit). J'en tire qu'une réponse bien tournée ne t'apprend rien sur sa justesse.

Sur le fait que la connaissance soit rangée dans le modèle : les auteurs du papier fondateur sur la génération assistée par documents l'écrivaient en 2020 - mettre à jour ces connaissances et retracer l'origine d'une réponse restaient « des problèmes de recherche ouverts » (Lewis et coll. (<https://arxiv.org/abs/2005.11401>), traduit). Et Kalai et coll. précisent que leur analyse vaut aussi pour les systèmes qui vont chercher des documents.

Pourquoi la même question donne deux réponses

Ce n'est pas une panne, c'est un choix de conception.

Une équipe l'a documenté en 2019 : prendre systématiquement le mot le plus probable produit un texte « plat, incohérent, ou qui tourne en boucle ». Leur méthode tire le mot suivant au sort dans la partie fiable de la distribution, et c'est ce tirage qui rend le texte lisible (Holtzman et coll., ICLR 2020 (<https://arxiv.org/abs/1904.09751>), traduit). C'est leur méthode ; je ne sais pas quel réglage utilise l'outil que tu ouvriras.

Ce que ça change pour toi :

- une réponse qui te plaît, garde-la : tu ne la retrouveras pas à l'identique ;
- ne bâtis pas de procédure qui suppose une réponse toujours identique ;
- une réponse décevante mérite une relance - c'est le fonctionnement normal.

Et sa limite : relancer change la formulation, ça ne vérifie rien. Deux fois la même réponse ne la rend pas vraie.

Le cas documenté qui doit te rester

La situation. Des chercheurs demandent à un modèle une date de naissance précise, en ajoutant explicitement : réponds seulement si tu sais.

Ce qu'ils obtiennent. Trois essais, trois dates différentes, toutes fausses, toutes bien formatées - « 03-07

», « 15-06 », « 01-01 » - d'un modèle libre parmi les plus avancés, en septembre 2025 (Kalai et coll., arXiv:2509.04664 (<https://arxiv.org/abs/2509.04664>), §1, traduit). N'attends pas de reproduire ces trois dates-là : le modèle testé n'est pas nommé dans le passage. Ce qui se reproduit, c'est le motif - de l'assurance sur un fait rare.

Le signal. Une précision élevée sur un fait rare, et une réponse qui bouge d'un essai à l'autre.

Le mécanisme. Une date de naissance peu citée n'offre aucune régularité. Les mêmes auteurs notent que les modèles se trompent rarement sur les faits très souvent repris. C'est ce qui rend l'erreur difficile à anticiper : elle ne se produit pas là où tu penserais à la tester.

Un second profil, qui ne donne pas ce signal

C'est le développement de la ligne du module sur les réponses stables. Un seuil réglementaire, un délai de garantie, un tarif conventionnel sont des formulations figées, écrites partout de la même façon, donc bien représentées dans ce que le modèle a lu.

Ce qui suit est un raisonnement à partir du mécanisme, pas une mesure : je n'ai pas compté. Une réponse de ce type a toutes les chances de tomber pareil d'un essai à l'autre, y compris quand elle reprend l'état du texte avant sa dernière réforme - le modèle n'a aucun moyen de savoir laquelle des deux versions est en vigueur.

Ici, la stabilité de la réponse ne prouve rien : elle dit seulement que la formulation était fréquente. Le seul contrôle est celui de la branche production - où est-ce publié, et est-ce que j'ouvre la page ?

Le même tri, sur trois autres métiers

- Assistante de direction, un compte-rendu à partir de ses notes : transformation. Les échéances sont-elles celles des notes ?
- Artisan, une réponse à un appel d'offres citant une norme : production. La référence s'ouvre-t-elle, et dit-elle ça ?
- Gérante, comparer deux contrats qu'elle fournit : transformation. La garantie citée dans le tableau est-elle bien celle de la clause, avec sa durée et ses exclusions ? Et si une garantie apparaît sans figurer dans aucun des deux contrats, c'est une production - elle se retire, elle ne se vérifie pas.

Deux mots, et c'est tout

Un modèle, c'est le composant entraîné qui produit le texte. Pas l'application dans laquelle tu tapes.

Le contexte, c'est ce que tu lui fournis pour orienter sa réponse : ta demande, la conversation, les documents joints. C'est le levier de « Bien lui parler, pour arrêter d'obtenir du tiède ».

Les sources de ce module sont consignées dans docs/f1-PROVENANCE.md`, avec la date de consultation et l'extrait constaté. Les liens ont été contrôlés le 06/08/2026.