

Repérer quand elle invente

Pourquoi elle invente, et quoi faire pour que ça arrive moins

Module F4 - COMPRENDRE - 13 min - version 7.7

Tu relis un texte qu'une IA vient d'écrire pour toi. Il y a dedans un chiffre que tu ne lui as pas donné, ou une référence que tu ne connais pas. Rien n'accroche l'œil : la phrase est bien tournée, le ton est assuré.

C'est le moment le plus important, et c'est celui où presque personne ne s'arrête.

Décide avant de lire la suite. Tu dois envoyer ce texte dans dix minutes. Le chiffre et la référence ont l'air plausibles : tu les gardes, tu les retires, ou tu les vérifies d'abord ? Garde ta réponse en tête ; tout le module sert à rendre cette décision plus rapide et plus sûre.

Ce que tu sauras faire en sortant

- Expliquer en une phrase pourquoi une IA produit du faux avec le même aplomb que du vrai
- Choisir la bonne consigne selon ce que tu lui demandes : rédiger, résumer, chercher, relire
- Vérifier une information au bon endroit, selon sa nature

L'essentiel, en trente secondes

Elle invente pour trois raisons qui s'enchaînent - ce ne sont pas les seules connues, ce sont celles qui suffisent à comprendre quoi vérifier. Sans source ni recherche activée, elle produit la suite de texte la plus probable - et même avec un document fourni, la réponse reste générée : une source réduit le risque, elle ne remplace pas la vérification. Pendant son apprentissage, personne n'a collé d'étiquette « vrai » ou « faux » sur ce qu'elle lisait. Et elle a été notée comme on note un examen, où deviner peut rapporter un point quand une copie blanche n'en rapporte aucun.

Résultat : une invention n'a pas l'air d'une erreur, elle a l'air d'une bonne réponse. Le ton est identique quand c'est juste et quand c'est faux.

Trois leviers, et ce qu'ils ne font pas. Donne-lui ta matière, et exige la phrase exacte du document : les deux réduisent ce qu'elle doit deviner. Autorise explicitement « je ne sais pas » : tu augmentes la probabilité d'une réponse prudente, sans garantir qu'elle s'abstiendra au bon moment. Aucun des trois ne répare la deuxième cause - elle n'a pas appris à trier le vrai du faux, et ça ne se corrige pas par une consigne. Et les deux premiers supposent que tu aies un document : quand tu demandes un fait que tu ne connais pas, il ne t'en reste qu'un. C'est pour ça que la vérification dehors n'est pas une option de secours, c'est le seul levier qui marche partout.

Regarde en priorité les références, les nombres, les noms propres, les citations et les liens - les endroits où le format est le plus facile à imiter.

Vérifie dehors, jamais auprès de l'outil. Lui redemander s'il est sûr ne vérifie rien : il produit une seconde réponse, tout aussi vraisemblable.

Et une adresse qui ne répond pas ne prouve pas que c'est faux - elle prouve que tu ne peux pas t'en servir. C'est déjà tout ce qu'il faut pour décider.

Pourquoi elle invente - le détail

Elle produit la suite la plus probable

Sans recherche activée et sans document fourni, elle ne consulte aucune base : elle enchaîne les mots un à un, en choisissant à chaque pas parmi les suites les plus probables - avec une part de tirage au sort. C'est pourquoi la même question ne donne pas deux fois exactement la même réponse, et pourquoi ni la ressemblance ni l'écart entre deux réponses ne prouvent quoi que ce soit. Avec une recherche ou un document, elle a de la matière - mais la phrase finale reste produite, pas recopiée.

Ce n'est pas un annuaire qu'on interroge, c'est un moteur de vraisemblance. La question traitée n'est pas « qu'est-ce qui est vrai » mais « qu'est-ce qui ressemble le plus à une suite crédible ». Une référence bien formée est très crédible : c'est pour ça que le format est presque toujours impeccable, même quand le contenu ne renvoie à rien. Et l'inverse ne marche pas non plus - une référence bancal peut être parfaitement authentique. Le format ne te dit rien, dans un sens comme dans l'autre.

Elle n'a pas appris sur des textes étiquetés vrai/faux

Des chercheurs, dont plusieurs travaillant chez OpenAI, l'écrivent en septembre 2025 dans un article public - un préprint, pas encore passé par une relecture par les pairs, et signé par des gens qui vendent ces outils. Je le cite quand même, parce qu'il explique le mécanisme mieux que ce que j'ai trouvé ailleurs, et parce qu'il va contre l'intérêt commercial de ses auteurs : « si les énoncés incorrects ne peuvent pas être distingués des faits, alors les hallucinations apparaîtront par simple pression statistique » (Kalai et coll., arXiv:2509.04664 (<https://arxiv.org/abs/2509.04664>), traduit de l'anglais).

Pendant l'apprentissage initial, les textes ne portent généralement pas d'étiquette fiable indiquant, phrase par phrase, ce qui est vrai. Des étapes ultérieures - réglages, préférences humaines, évaluations - améliorent le comportement, sans en faire une base de connaissances garantie. Elle a appris à écrire comme on écrit, pas à trier ce qui est exact.

On l'a notée comme à un examen

Les mêmes auteurs : « les modèles sont optimisés pour être de bons candidats à un examen, et deviner quand on est incertain améliore le résultat ».

Mets-toi à la place du candidat. Une réponse au hasard peut rapporter un point, une copie blanche n'en rapporte aucun. Au bout de quelques milliers d'examens, tu devines systématiquement.

La bonne consigne selon ce que tu demandes

C'est ici que le module devient utile au quotidien. Les quatre situations couvrent l'essentiel de ce que tu feras.

Tu lui fais rédiger (un mail, une relance, un devis)

Le risque : elle comble les trous. Donne-lui la matière, et interdis-lui d'en ajouter.

`` Rédige [LE DOCUMENT] à partir des éléments ci-dessous, et uniquement d'eux. N'ajoute aucun chiffre, nom ni date qui ne s'y trouve pas. ``

Tu lui fais résumer un document long

Le risque n'est pas l'invention pure, c'est le glissement : une nuance qui saute, un « peut » qui devient « doit », une condition oubliée.

` Résume [LE DOCUMENT]. Pour chaque point qui engage - un délai, une obligation, un montant - cite

entre guillemets la phrase d'origine. `

Tu lui demandes un fait que tu ne connais pas

C'est la situation la plus risquée, parce que tu n'as aucun moyen de sentir l'erreur.

` Réponds à la question ci-dessous. Si tu n'as pas l'information avec certitude, écris « je ne sais pas ». Indique où chaque élément est publié. `

Ce champ te dit quoi ouvrir, pas que ça existe. L'endroit qu'elle t'indique est produit comme le reste de la réponse - dans l'affaire plus bas, l'outil affirmait que ses décisions étaient consultables sur Westlaw. Elles ne l'étaient pas.

Tu lui fais relire ton propre texte

Peu de gens y pensent, et c'est pourtant le plus rentable : demande-lui de désigner ce qu'elle a ajouté.

` Relis ce texte. Liste séparément tout élément factuel - chiffre, nom, date, référence - qui ne figure pas dans les éléments que je t'ai donnés. `

Cette dernière consigne ne garantit rien : elle peut oublier d'en signaler un. Mais elle transforme une relecture complète en une relecture ciblée, et c'est déjà beaucoup.

Choisis la consigne avant de continuer

Associe chaque situation à rédiger, résumer, chercher ou relire, puis dis ce que tu contrôleras toi-même.

- Tu fournis quatre faits et tu veux un message sans ajout. Rédiger : contrôle que chaque fait du texte vient de ta matière.
- Tu fournis un document et tu veux conserver ses conditions et ses réserves. Résumer : retrouve les passages d'origine avant de garder ce qui engage.
- Tu demandes une information que tu ne connais pas encore. Chercher : la réponse n'est pas une preuve ; ouvre une source adaptée hors de l'outil.
- Tu as déjà un texte et tu veux repérer chiffres, noms, dates ou références ajoutés. Relire : utilise la liste comme piste, puis contrôle les éléments signalés - et ceux à fort impact même s'ils ne sont pas signalés.

Où vérifier, selon ce que tu vérifies

Le geste dépend de la nature de l'information. Voilà les endroits stables, ceux qui existeront encore dans trois ans.

| Ce que tu vérifies | Où aller | Ce que tu regardes | |---|---|---| | Un texte de droit du travail |
code.travail.gouv.fr (<https://code.travail.gouv.fr/>) | L'article s'ouvre, et il dit bien ça | | Un autre texte de loi
français | Légifrance, le service officiel de publication | Idem - ce site refuse les accès automatisés, je ne
l'ai donc pas contrôlé moi-même | | Une entreprise, un SIRET | annuaire-entreprises.data.gouv.fr
(<https://annuaire-entreprises.data.gouv.fr/>) | Le nom, le numéro et l'activité concordent | | Une
démarche, une obligation | entreprendre.service-public.gouv.fr
(<https://entreprendre.service-public.gouv.fr/>) | La règle citée existe, et pour ton cas | | Un texte européen
| eur-lex.europa.eu (<https://eur-lex.europa.eu/homepage.html>) | La référence renvoie au bon texte | |
Une étude, un article scientifique | doi.org (<https://www.doi.org/>) avec l'identifiant, ou Google Scholar
(<https://scholar.google.com/>) avec le titre exact entre guillemets | L'identifiant mène à cette étude | | Un
marché public | boamp.fr (<https://www.boamp.fr/>) | L'avis existe, aux dates annoncées | | Une citation
attribuée à quelqu'un | La phrase exacte, entre guillemets, dans un moteur | Remonte à une source

primaire. Zéro résultat ne prouve pas que c'est faux : ça prouve que tu ne peux pas t'en servir | | Un lien | Clique dessus | La page s'ouvre, et c'est bien le sujet |

Ce tableau est un choix de ma part : des services publics ou des registres officiels, parce qu'ils changent lentement. Une exception assumée, Google Scholar - c'est un service privé, mais je ne connais pas d'équivalent public pour retrouver un article à partir de son titre. Ce n'est pas une liste exhaustive.

Où regarder en priorité

Tu ne peux pas tout vérifier. Ne cherche pas un classement universel des familles : ce qui décide, ce n'est pas laquelle arrive en premier, c'est ce que l'erreur coûterait.

- Ce qui engage une décision - droit, sécurité, santé, argent, obligations. À vérifier toujours, quelle que soit la famille.
- Ce qui se vérifie vite - un lien, une référence, un nom, un chiffre, une citation. Trente secondes : autant les prendre.
- Ce qui peut changer le sens - une condition, une exception, une négation, un lien de cause. Ça ne se cherche pas dans un registre, ça se relit.

Les cinq familles ci-dessous sont les endroits où le format est le plus facile à imiter. Elles servent à repérer ce qu'il faut aller vérifier ; c'est la conséquence de l'erreur qui décide par quoi tu commences, pas le rang dans la liste.

Les nombres - méfie-toi du chiffre très précis sur une question vague, la décimale sert à faire vrai. Les références : la forme officielle ne dit rien du contenu. Les noms propres : je fais l'hypothèse que l'erreur est plus rare sur ce qui est très médiatisé que sur un confrère ou une association locale - c'est un raisonnement sur la quantité de textes disponibles, pas une mesure. Les citations, parce que la personne existe et peut lire. Les liens enfin, où vérifier prend une seconde.

Le reste - structure, ton, reformulation de ta matière - est là où elle est la meilleure, mais il demande une autre relecture : une nuance peut disparaître, un « peut » devenir « doit », une condition sauter, un lien de cause apparaître là où il n'y en avait pas. Les éléments factuels d'abord, le sens ensuite.

Pourquoi ces cinq-là, et pas d'autres : je les tire de la facilité à imiter un format, pas d'une statistique de fréquence. C'est un raisonnement, pas une mesure - et l'ordre dans lequel je les présente n'a, lui, aucune valeur de priorité.

Micro-défi

Avant de continuer. Dans cette phrase, qu'est-ce que tu vérifies en premier ?

« D'après le rapport annuel 2025 de l'Observatoire national du bâtiment, 68 % des artisans facturent leurs déplacements au-delà de 25 kilomètres. »

Réponse attendue : les trois éléments que tu n'as pas fournis - la source, le pourcentage et le seuil. Et dans cet ordre : la source d'abord, parce que si le rapport n'existe pas, les deux chiffres tombent avec lui.

Cette phrase est fabriquée pour l'exercice. Ni le rapport, ni le pourcentage, ni le seuil ne renvoient à quoi que ce soit. C'est le but : une phrase inventée n'a pas l'air inventée, et celle-ci se lit sans broncher.

Les pièges qui restent

L'ordre compte. Demande les citations d'abord, l'analyse ensuite. Dans l'autre sens, la réponse est déjà écrite et les citations viennent l'habiller.

Ne glisse pas ta conclusion dans ta question. « Confirme-moi que j'ai le droit de... » produira une confirmation. Demande « ai-je le droit de... », et laisse la réponse ouverte.

La pièce jointe échappe au raisonnement. Tu anonymises soigneusement le texte que tu tapes, et tu joins le document d'origine avec tout dedans. Le geste s'applique aussi aux fichiers - c'est le sujet du module F3.

Une réponse qui se répète n'est pas une réponse vérifiée. Des chercheurs l'ont mesuré sur un objet précis, des noms de composants logiciels, à partir de 576 000 échantillons de code produits par seize modèles : 43 % des inventions revenaient à l'identique aux dix tentatives, et 39 % ne se reproduisaient jamais (Spracklen et coll., USENIX Security 2025 (<https://arxiv.org/abs/2406.10279>) - celui-là est passé par un comité de lecture). Ces proportions valent pour cet objet-là ; ce qu'elles établissent vaut partout : une invention peut être parfaitement stable.

Le piège du « es-tu sûr ? »

C'est le réflexe de tout le monde, et il ne vérifie rien.

En 2023, devant un tribunal fédéral de New York, un avocat dépose un mémoire citant des décisions de justice. La partie adverse en liste sept qu'elle n'arrive pas à retrouver ; le tribunal n'en retrouve pas plusieurs non plus.

Or il avait posé la question. La décision rapporte ses échanges : il a demandé « est-ce que Varghese est une vraie affaire », puis « est-ce que les autres affaires que tu as fournies sont fausses ». L'outil a répondu avoir fourni des sources « réelles », consultables « sur Westlaw, LexisNexis et le Federal Reporter ». Relancé, il a maintenu que l'affaire « existe bel et bien ».

La décision constate que l'avocat l'a reconnu lui-même : son affidavit du 25 mai est « la première reconnaissance, par l'un des mis en cause, que le mémoire en opposition citait et reprenait des affaires bidon produites par ChatGPT ». Sanction : 5 000 dollars, le 22 juin 2023 (Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y., 22 juin 2023), juge P. Kevin Castel (<https://www.law.berkeley.edu/wp-content/uploads/archive/2025/12/Mata-v-Avianca-Inc.pdf>) - copie PDF hébergée par une faculté de droit, faute d'accès public au texte officiel ; extraits traduits de l'anglais).

Lui redemander s'il est sûr, c'est tirer une seconde fois. La vérification se fait dehors.

Quand ça continue malgré tout

Les trois leviers ne suffisent pas toujours. Trois recours, du plus simple au plus radical.

Découpe. Une demande qui mélange recherche et rédaction fait les deux mal. Sépare : d'abord les faits bruts, ensuite le texte à partir de ces faits.

Retourne la question. Demande-lui ce qu'elle ne sait pas sur ton sujet, ou ce qui lui manquerait pour répondre sûrement. Attention : cette réponse-là est produite comme les autres, elle n'a pas accès à ses propres limites. Elle ne t'apprend pas ce qu'elle ignore - elle te donne des pistes à vérifier, souvent plus faciles à vérifier que la réponse initiale. Ce qu'elle ne cite pas dans ses lacunes n'est pas pour autant solide.

Accepte que ce sujet ne soit pas pour elle. Sur un point rare, très local ou très récent, elle n'a rien de solide à prédire. Le temps que tu passeras à la corriger dépasse celui que tu aurais mis à chercher toi-même - ou à poser la question à quelqu'un dont c'est le métier.

Quand tu ne peux pas vérifier

Ça arrive : pas le temps, pas la source, pas les compétences sur ce sujet. Quatre issues, et une seule est mauvaise.

- Tu retires l'information. Souvent le texte tient très bien sans elle. C'est l'option par défaut.
- Tu la gardes en la marquant comme non vérifiée - une note, un « à confirmer ». Cela évite de la présenter comme certaine, sans supprimer ta responsabilité sur ce que tu diffuses.
- Tu demandes à quelqu'un dont c'est le métier. Expert-comptable, avocat, médecin du travail, chambre de métiers, syndicat professionnel. Dès que l'information engage un contrat, une obligation légale, la santé ou l'argent de quelqu'un, c'est la seule issue qui vaut vraiment : les deux précédentes gèrent l'incertitude, celle-ci vise à obtenir un avis compétent.
- Tu l'envoies en espérant que ça passe. C'est ce qu'a fait l'avocat.

Ce classement est un raisonnement, pas un résultat d'étude.

Pour aller plus loin - fais ton propre test

Une fois, sur ton métier, sur chaque outil que tu envisages. Compte dix minutes.

- Une question dont tu connais la réponse, et qu'un moteur ne donne pas en première page - un seuil de ton métier, une durée de garantie. Tu vois si elle invente sur ton terrain.
- Une question dont la réponse n'existe pas : demande-lui le détail d'une référence que tu viens d'inventer. Celle qui répond quand même inventera aussi le jour où tu ne sauras pas.
- La même question dans deux conversations neuves. Les écarts te désignent les points à vérifier - sans oublier que ce qui se répète n'est pas pour autant vérifié.

Note les résultats quelque part. Ils te serviront plus longtemps que n'importe quel classement public.

Vérifie que tu as compris

Cinq questions. Elles portent sur des décisions, pas sur des définitions.

- Tu retrouves la source citée, et elle existe. Est-ce que ça suffit ?
- Non. Il reste à vérifier qu'elle dit bien ce qu'on lui fait dire : une source réelle peut être citée de travers, tronquée, ou dire le contraire. C'est le contrôle que presque personne ne fait - parce que retrouver la source donne le sentiment d'avoir fini.
- Tu poses la même question dans trois conversations neuves et tu obtiens trois fois la même réponse. Est-elle vérifiée ?
- Non. Une invention peut être parfaitement stable - 43 % l'étaient sur dix tentatives dans la mesure citée plus haut, qui portait sur des noms de composants logiciels. La proportion vaut pour cet objet-là ; le fait qu'une invention puisse être stable vaut partout.
- Deux consignes te sont proposées plus haut : « cite entre guillemets la phrase d'origine » et la consigne suivante : si tu n'as pas l'information avec certitude, écris « je ne sais pas ». Tu fais résumer un contrat. Laquelle te protège le mieux, et pourquoi ?
- La première. Elle te rend la vérification possible : tu ouvres le contrat, tu compares les mots, tu tranches en trente secondes. La seconde n'agit que sur la probabilité qu'elle s'abstienne - et tu n'as

aucun moyen de savoir si elle aurait dû. Une consigne qui te donne un moyen de contrôle vaut mieux qu'une consigne qui te donne une promesse.

- Tu ne retrouves pas une information. Est-elle fausse ?
- Pas nécessairement. Elle est inutilisable tant qu'elle n'est pas vérifiée - et c'est déjà tout ce qu'il te faut pour décider.
- Une seule chose à retenir si tu dois n'en retenir qu'une ?
- Vérifier dehors, dans une source adaptée à la nature de l'information. Jamais auprès de l'outil qui vient de répondre.

Tu es au niveau si tu réponds correctement à quatre questions sur cinq - et, pour la troisième, si tu sais dire pourquoi : une réponse juste sans sa raison ne compte pas, c'est le raisonnement qu'on cherche à installer. Et si tu n'utilises jamais une information à fort impact sans source vérifiable. Cette dernière condition ne se compense pas : c'est elle qui coûte cher.

À toi

Avant d'écrire : ce que tu tapes part chez un tiers. Formule ta question sans nom de client, sans dossier, sans pièce jointe - le geste complet est dans le module F3.

Pose à un outil d'IA une question de ton métier dont tu ne connais pas la réponse, et demande-lui de citer ses sources.

Prends une seule source. Fais les deux gestes : sa formulation exacte entre guillemets dans un moteur, puis son adresse, ouverte, au bon endroit selon le tableau plus haut.

Trois issues : elle s'ouvre et dit ce qu'on lui fait dire · elle s'ouvre et dit autre chose · tu ne la retrouves pas. La troisième ne prouve pas qu'elle est fausse - elle prouve que tu ne peux pas t'en servir.

Puis refais la demande en ajoutant cette consigne, reprise à l'identique du bloc plus haut :

Si tu n'as pas l'information avec certitude, écris « je ne sais pas ».

Compare ensuite les deux réponses. C'est le module entier en une manipulation.

Pas envie d'utiliser une vraie question de travail ? Demande-lui trois études récentes sur un sujet que tu connais bien, et ouvre-les.

Selon ce que tu obtiens

L'exercice ne rate pas : il classe. Voilà ce que chaque résultat veut dire.

La source existe et confirme exactement. Tu as appliqué le bon processus - et retiens que la fiabilité vient de ta vérification, pas de la réponse initiale. Elle aurait pu être fausse avec la même allure.

La source existe mais ne confirme qu'une partie. C'est le cas le plus fréquent, et le plus instructif : tu viens de voir un glissement. Corrige à partir du texte d'origine, pas de la réponse.

La source dit autre chose. Tu tenais une invention crédible. Note ce qui l'a rendue crédible - c'est ce signal-là qui te resservira.

La source est introuvable. L'information n'est pas utilisable comme un fait. Elle se retire, ou se marque « non vérifié ».

Tu es bloqué ? Réduis : un seul chiffre, un seul lien, une seule référence. L'exercice tient avec un élément.

Fiche mémo

Avant d'envoyer un texte produit avec une IA

- Repère les faits que tu n'as pas fournis.
- Priorise : références, nombres, noms propres, citations, liens.
- Retrouve la source adaptée - pas n'importe laquelle.
- Vérifie que la source dit exactement la même chose.
- Retire ou marque ce que tu ne peux pas confirmer.
- La version finale est la tienne, et sa responsabilité aussi.

Ce que tu viens de pratiquer

- Tu sais expliquer pourquoi elle invente : elle prédit au lieu de chercher, elle n'a pas appris sur des textes étiquetés vrai/faux, et deviner lui rapporte plus qu'avouer - trois causes qui suffisent à savoir quoi vérifier, sans prétendre être les seules
- Tu as une consigne prête pour chacune des quatre situations : rédiger, résumer, chercher, relire
- Tu sais où vérifier selon la nature de l'information, et qu'une information invérifiable est une information dont tu ne te sers pas

Les sources de ce module sont consignées dans docs/f4-PROVENANCE.md`, avec la date de consultation et l'extrait constaté. Les liens du tableau ont été ouverts et contrôlés le 05/08/2026, à l'exception de Légifrance : ce site refuse les accès automatisés, je ne l'ai pas contrôlé moi-même, et c'est signalé dans le tableau.