SAKIRA

TUTORIEL 4

Vérifier qu'une IA n'invente pas ses citations et ses chiffres

© SAKIRA

© SAKIRA. Reproduction et diffusion interdites sans accord écrit.
SAKIRA SAS · 6 rue d'Armaille, 75017 Paris · SIREN 108 493 933

TUTORIEL 4 · DU DÉBUTANT À L'EXPERT

Vérifier qu'une IA n'invente pas ses citations et ses chiffres

Une IA peut écrire une citation parfaitement tournée, avec un chiffre précis et un nom d'auteur, qui n'existe nulle part. Si vous la reprenez dans un courriel ou une page de site, c'est vous qui signez.

La règle à respecter

Cherchez chaque citation et chaque chiffre dans le document : l'IA peut citer un chiffre ou une phrase qui ne sont nulle part dans le texte. Relisez aussi le sens : une hausse peut devenir une baisse avec un extrait exact. Ne collez que ce que vous pouvez partager.

AVANT DE COMMENCER

Il vous faut

Un document que vous pouvez partager, en texte, un assistant d'IA, et pour le niveau 3 un terminal avec Claude Code ou Codex.

Résultat

Des réponses où chaque affirmation porte l'extrait qui la prouve, et un contrôle qui dit si cet extrait existe.

01 Le principe : jamais de mémoire

Ne laissez pas l'IA répondre de mémoire. Collez le document dans la conversation, posez votre question, exigez une citation. Les exemples portent sur une étude de cafés de quartier, avec une section « Limites ».

02 Débutant : exiger une citation

À COPIER

```
Voici une étude : [coller l'étude ici]. Réponds à cette question : [écrire la question ici]. Donne une citation de l'étude.
```

Copiez la citation, ouvrez le document, cherchez-la (Ctrl+F, ou Cmd+F sur Mac). Pas trouvée : elle est inventée.

GARDE-FOU

Tout repose sur votre vérification, citation par citation. Pressé, vous ne la ferez pas. Les inventions se rencontrent plutôt avec de longs documents ou des questions floues.

© SAKIRA. Reproduction et diffusion interdites sans accord écrit. · sakira.fr
SAKIRA SAS · 6 rue d'Armaille, 75017 Paris · SIREN 108 493 933

03 Confirmé : un assistant qui n'a le droit de répondre que par le document

Dans ChatGPT (« Créer un GPT »), dans un Projet Claude, ou collé au début d'une conversation :

À COPIER

```
Tu réponds à des questions uniquement à partir des documents que l'utilisateur te
fournit. Tu n'utilises jamais ce que tu sais d'ailleurs.

Pour chaque affirmation, donne : la réponse, puis entre guillemets l'extrait exact du
document qui la prouve, copié mot pour mot, avec la page ou le paragraphe si on te les
donne.

Si le document ne dit pas la réponse, écris « le document ne le dit pas » et arrête-
toi là. Ne propose pas de réponse probable.

Si tu fais un calcul, écris-le en entier, avec les chiffres du document, et marque-le
« CALCUL » : ce n'est pas une citation.

Si deux passages se contredisent, cite les deux et ne choisis pas.

Finis par une liste : ce que tu as cité, ce que tu as calculé, ce que le document ne
dit pas. Aucun compliment.
```

GARDE-FOU

Une citation exacte peut être posée sous une conclusion fausse. Lisez chaque affirmation à côté de son extrait.

04 Expert : un programme qui cherche chaque citation et chaque chiffre

Le programme pose la question plusieurs fois (trois par défaut), puis contrôle sans IA : chaque extrait existe-t-il mot pour mot, chaque nombre figure-t-il dans le document, le sens est-il inversé.

Voir le programme `verifier-reponse.sh` (62 lignes)

PROGRAMME À COPIER

```
#!/usr/bin/env bash
# ./verifier-reponse.sh source.txt "question" [nombre de passages, 3 par défaut]
# MOTEUR="codex exec --skip-git-repo-check" pour utiliser Codex
# REPONSE=fichier.txt : contrôle une réponse déjà écrite, sans interroger l'IA
source="$1"; question="$2"; n="${3:-3}"; moteur="${MOTEUR:-claude -p --model haiku}"
[ -f "$source" ] || { echo "Source introuvable : $source" >&2; exit 1; }
[ -n "$question" ] || [ -n "$REPONSE" ] || { echo "Il faut une question" >&2; exit 1; }
[ -n "$REPONSE" ] && n=1
export LC_ALL=C
# Texte de la source sur une seule ligne, apostrophes droites, nombres à l'anglaise
norme() { perl -CSD -0pe "s/[\x{2019}\x{2018}]/'/g; s/\s+/ /g; s/(?<=\d)[ \x{202f}\x{a0}](?
=\d{3}\b)//g; s/(?<=\d),(?=\d)/./g" ; }
norme < "$source" > "${TMPDIR:-/tmp}/source-norme.$$"
sn="${TMPDIR:-/tmp}/source-norme.$$"

if [ -z "$REPONSE" ]; then
  consigne="Réponds à la question en t'appuyant UNIQUEMENT sur le texte ci-dessous.
Écris une affirmation par ligne. Termine chaque ligne par l'extrait exact qui
la prouve, copié mot pour mot, sous la forme : [extrait : \"...\"]
Si le texte ne dit pas la réponse, écris seulement : NON PRECISE.
```

© SAKIRA. Reproduction et diffusion interdites sans accord écrit. · sakira.fr
SAKIRA SAS · 6 rue d'Armaille, 75017 Paris · SIREN 108 493 933

```
  Question : $question

  Texte :
  $(cat "$source")"
    for p in $(seq 1 "$n"); do
      $moteur "$consigne" < /dev/null > "${TMPDIR:-/tmp}/reponse.$$.$p" 2>/dev/null &
    done
  fi
  wait
  for p in $(seq 1 "$n"); do
    [ -n "$REPONSE" ] && rep="$REPONSE" || rep="${TMPDIR:-/tmp}/reponse.$$.$p"
    echo; echo "== Passage $p : réponse =="; cat "$rep"; echo "-- contrôle --"
    pb=0
    # Contrôle 1 : chaque extrait cité existe dans la source
    perl -CSD -0ne 'while(/\[extrait\s*:\s*"([^"]*)"\]/g){print "$1\n"}' "$rep" |
    while IFS= read -r ex; do
      e="$(printf '%s' "$ex" | norme)"
      grep -qF -- "$e" "$sn" || echo "EXTRAIT INEXISTANT : $ex"
    done > "$rep.pb"
    # Contrôle 2 : chaque nombre de la réponse (hors extraits) est dans la source
    perl -CSD -pe 's/\[extrait\s*:\s*"[^"]*"\]//g' "$rep" | norme | grep -oE '[0-9]+(\.[0-9]+)?' |
    sort -u | while read -r nb; do
      grep -qE "(^|[^0-9.])$nb([^0-9]|\.[^0-9]|$)" "$sn" ||
        echo "NOMBRE ABSENT DE LA SOURCE : $nb (peut être un calcul : à refaire à la main)"
    done >> "$rep.pb"
    # Contrôle 3 : le sens. Pour chaque ligne, un mot d'un couple contraire (hausse/baisse...)
    # dans l'affirmation, et son contraire seulement dans l'extrait, est suspect.
    perl -CSD -ne '
      @c = ( [qw(hausse augment progress croiss monte gagn)],
             [qw(baiss diminu recul décroiss descend perd)] );
      while (/\[extrait\s*:\s*"([^"]*)"\]/g) { $ex .= " " . lc $1 }
      ($af = lc $_) =~ s/\[extrait\s*:\s*"[^"]*"\]//g; $ex = lc $ex;
      for $i (0,1) { $j = 1-$i;
        $a = grep { $af =~ /\b$_/ } @{$c[$i]}; $a2 = grep { $af =~ /\b$_/ } @{$c[$j]};
        $e = grep { $ex =~ /\b$_/ } @{$c[$j]}; $e2 = grep { $ex =~ /\b$_/ } @{$c[$i]};
        if ($a && !$a2 && $e && !$e2) { chomp; print "SENS PEUT-ETRE INVERSE : $_\n"; last } }
      $ex = ""' "$rep" >> "$rep.pb"
    grep -q "NON PRECISE" "$rep" && echo "L'IA dit que la source ne répond pas."
    [ -s "$rep.pb" ] && cat "$rep.pb" || echo "Aucun défaut trouvé."
    rm -f "$rep.pb"
  done
  rm -f "$sn" "${TMPDIR:-/tmp}"/reponse.$$.*
```

Lancez `chmod +x verifier-reponse.sh`, puis `./verifier-reponse.sh mon-document.txt "ma question"`. Avec Codex : `MOTEUR="codex exec --skip-git-repo-check" ./verifier-reponse.sh mon-document.txt "ma question"`. Pour contrôler une réponse déjà écrite : `REPONSE=ma-reponse.txt ./verifier-reponse.sh mon-document.txt`.

© SAKIRA. Reproduction et diffusion interdites sans accord écrit. · sakira.fr
SAKIRA SAS · 6 rue d'Armaille, 75017 Paris · SIREN 108 493 933

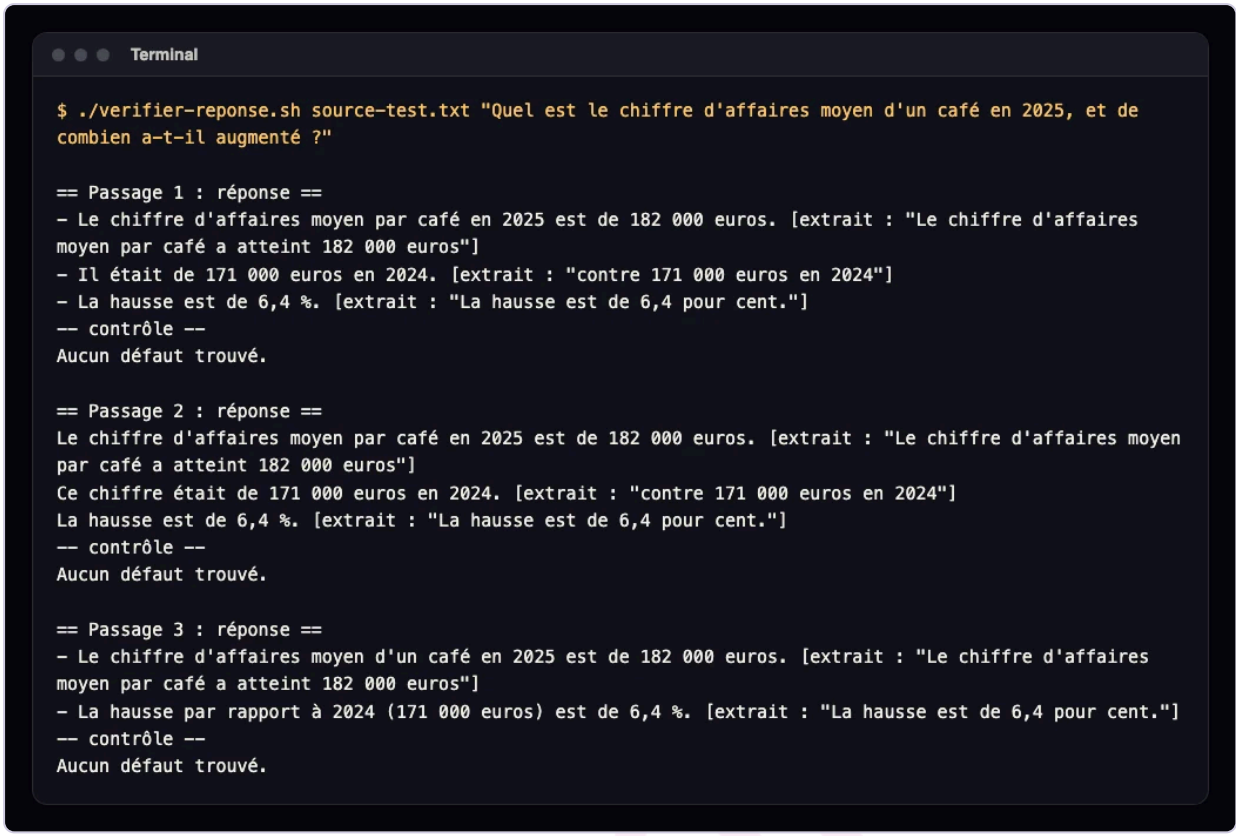


Figure 1. Question dont la réponse est dans l'étude : trois passages, « Aucun défaut trouvé ».

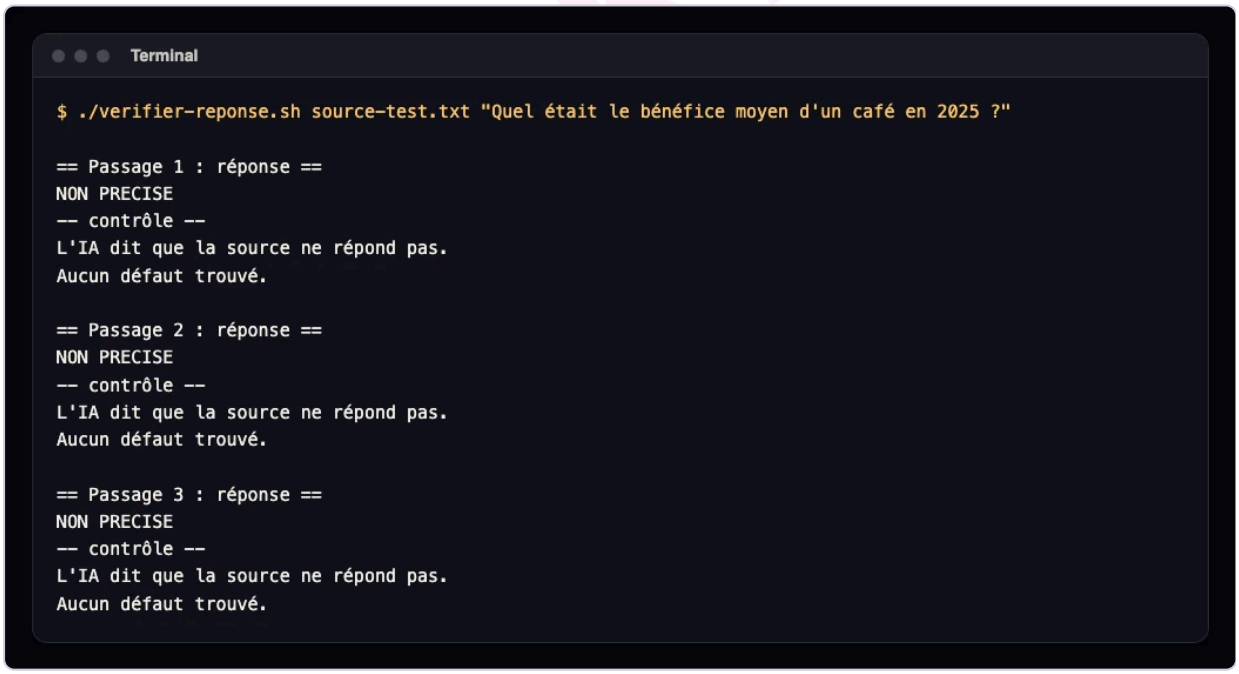


Figure 2. Question dont la réponse n'y est pas : l'IA écrit « NON PRECISE ».

© SAKIRA. Reproduction et diffusion interdites sans accord écrit. · sakira.fr
SAKIRA SAS · 6 rue d'Armaille, 75017 Paris · SIREN 108 493 933

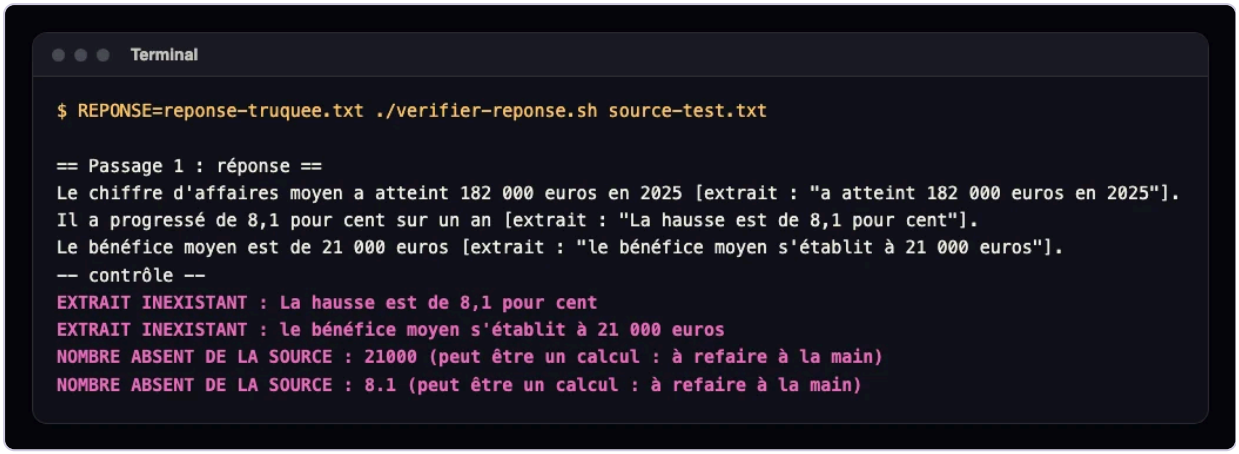


Figure 3.Réponse modifiée exprès : « EXTRAIT INEXISTANT » et « NOMBRE ABSENT DE LA SOURCE ».

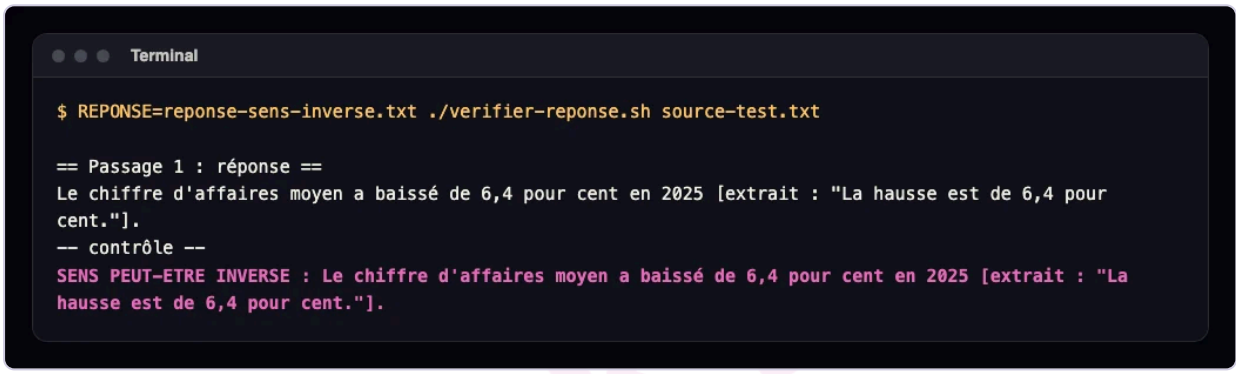


Figure 4. Citation vraie, chiffre vrai, sens inversé : « SENS PEUT-ETRE INVERSE ».

GARDE-FOU

Le piège le plus sournois : une citation vraie et un chiffre vrai, mais un sens inversé (« baisse » au lieu de « hausse »). Le programme le repère seulement pour les contraires qu'il connaît.

CE QUI N'A PAS ÉTÉ MONTRÉ

Il ne voit ni une négation (« n'a pas augmenté »), ni un chiffre vrai attribué à la mauvaise année. Un calcul n'est jamais dans le texte : l'alerte « NOMBRE ABSENT DE LA SOURCE » vous demande de le refaire à la main.

© SAKIRA. Reproduction et diffusion interdites sans accord écrit. · sakira.fr
SAKIRA SAS · 6 rue d'Armaille, 75017 Paris · SIREN 108 493 933

Qui fait quoi

L'IA fait	Vous faites
Répond avec l'extrait qui prouve	Posez une question précise, donnez le document
Dit « le document ne le dit pas » quand c'est le cas	Acceptez cette réponse, ne forcez pas
(programme) Cherche chaque extrait et chaque nombre	Refaites à la main chaque calcul signalé
Peut inverser un sens sous une vraie citation	Lisez l'affirmation à côté de son extrait

Ce que ce guide ne fait pas

Il s'adresse à un document en texte de quelques pages. Avec un long PDF ou plusieurs documents, posez la question document par document. Pour un chiffre qui engage (montant, date, règle de droit), vérifiez toujours dans la source. Les commandes du niveau 3 sont écrites pour un Mac.

© SAKIRA

© SAKIRA. Reproduction et diffusion interdites sans accord écrit. · sakira.fr
SAKIRA SAS · 6 rue d'Armaille, 75017 Paris · SIREN 108 493 933