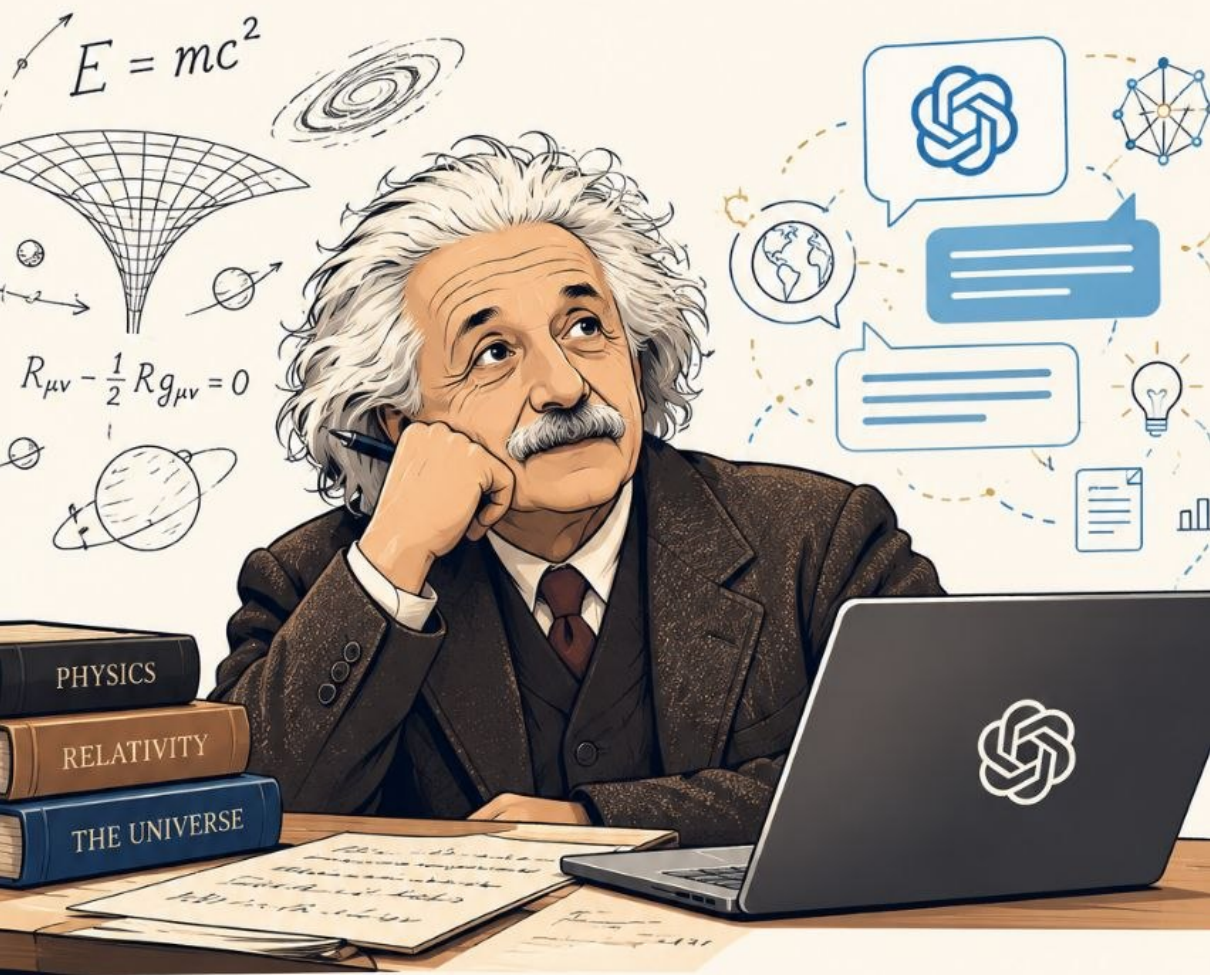


SI EINSTEIN AVAIT EU CHATGPT

Ce que l'intelligence artificielle pourrait
apprendre de notre manière de penser —
et ce que nous sommes en train de lui apprendre du monde



FRANCK MEVENGUE

PAGE DE TITRE INTÉRIEURE

SI EINSTEIN AVAIT EU CHATGPT

Ce que l'intelligence artificielle pourrait apprendre
de notre manière de penser — et ce que nous sommes
en train de lui apprendre du monde

FRANCK MEVENGUE

Note de transparence

Un livre sur l'IA, écrit avec l'IA

Avant toute chose, une petite expérience s'impose : ce livre parle d'intelligence artificielle... et il a lui-même été coécrit avec ChatGPT. Il aurait été dommage de parler de l'expérience sans la tenter.

À l'origine, ce projet est aussi une expérience éditoriale : explorer jusqu'où l'intelligence artificielle peut nous aider à concevoir, rechercher, structurer et rédiger un livre ou un article pertinent, sans renoncer à l'exigence intellectuelle ni à la créativité humaine.

Les idées défendues, les points de vue et les orientations de cet ouvrage relèvent de ma responsabilité d'auteur. ChatGPT m'a accompagné dans la recherche documentaire, l'exploration des idées, la structuration du manuscrit et le travail de rédaction. Les sources et les affirmations importantes ont vocation à être examinées et vérifiées par mes soins. Une réponse convaincante de l'IA ne constitue jamais, à elle seule, une preuve.

Je tiens à rendre cette collaboration transparente. Nous acceptons de plus en plus que l'IA participe à la conception de logiciels, à l'analyse de données ou à la productivité des entreprises. Pourquoi l'art, l'écriture et la création devraient-ils être condamnés à rester à l'écart de cette exploration ?

L'outil ne remplace pas nécessairement l'auteur : il peut aussi déplacer les frontières de ce qu'il est capable d'imaginer, d'explorer et de réaliser. Encore faut-il savoir lui poser les bonnes questions, exercer son jugement et assumer les choix finaux.

Après tout, si Einstein avait eu ChatGPT, il aurait peut-être testé ses hypothèses avec une intelligence artificielle. Pour ma part, j'ai commencé par tester si elle pouvait m'aider à écrire un livre sur ce sujet.

La boucle est bouclée. Ou presque : il reste encore à vérifier les sources.

PLAN DE L'OUVRAGE

Ce que le livre traverse

Prologue, dix parties et épilogue

Les distinctions entre faits, inférences, hypothèses et expériences de pensée structurent la lecture. Les identifiants entre crochets renvoient au registre scientifique du projet.

EXPÉRIENCE DE PENSÉE

2026 : Einstein ouvre ChatGPT

PEUT-ON MODÉLISER UNE PENSÉE ?

- 1 · Qu'est-ce qu'une pensée ?
- 2 · L'esprit laisse des traces
- 3 · Peut-on connaître quelqu'un sans être lui ?

SI EINSTEIN AVAIT EU CHATGPT

- 4 · Einstein en 2026
- 5 · Le chatbot personnel d'Einstein
- 6 · Le modèle d'Einstein
- 7 · Einstein aurait-il reconnu son propre modèle ?

COMMENT UNE IA APPREND

- 8 · Des données aux représentations
- 9 · Les modèles de fondation
- 10 · ChatGPT n'est pas une base de données
- 11 · Le contexte, la mémoire et les paramètres

L'HUMAIN COMME DONNÉE

- 12 · Chaque conversation est-elle une donnée ?
- 13 · Le chercheur nourrit-il l'IA ?
- 14 · L'IA apprend-elle aussi nos méthodes ?

L'IA ET LE MODÈLE DU MONDE

- 15 · Qu'est-ce qu'un « modèle du monde » ?
- 16 · L'IA possède-t-elle une représentation du monde ?
- 17 · De millions de traces à une représentation collective

UNE IA QUI GRANDIT AVEC L'HUMANITÉ

- 18 · 2026 → 2030 → 2040 → 2050
- 19 · Chaque génération hérite-t-elle de la précédente ?

20 · La boucle IA - humanité

LES MODÈLES HUMAINS

21 · L'IA peut-elle apprendre « qui nous sommes » ?

22 · Le jumeau intellectuel

23 · Une bibliothèque de personnalités scientifiques

LE PROBLÈME DE LA VÉRITÉ

24 · Données ≠ pensée

25 · Corrélation ≠ compréhension

26 · Hallucination

27 · Le biais du miroir

LE GRAND PARADOXE

28 · Nous apprenons à l'IA à nous comprendre

29 · Mais l'IA apprend aussi nos modèles du monde

30 · Qui apprend de qui ?

2050 : QU'AVONS-NOUS CRÉÉ ?

31 · La mémoire artificielle de l'humanité

32 · L'intelligence collective artificielle

33 · Si Einstein avait vraiment eu ChatGPT

ÉPILOGUE

Le miroir

2026 : Einstein ouvre ChatGPT

Une scène impossible pour poser une question réelle : que peut apprendre une machine des traces d'une pensée ?

■ EXPÉRIENCE DE PENSÉE

Tout le récit de ce prologue est fictif. Albert Einstein est mort en 1955. Les gestes, l'écran, les questions et les réactions décrits ici n'ont aucune valeur historique. Ils servent uniquement à construire une expérience de pensée.

Une expérience de pensée organise un scénario imaginé afin d'explorer les conséquences d'une question ou de principes. Elle ne transforme pas ce scénario en événement historique [\[SRC-038\]](#).

La fenêtre blanche

Imaginons une pièce de travail en 2026. Sur la table, des feuilles portent des équations, des flèches, des mots repris deux fois. Un carnet est ouvert près d'un ordinateur. Albert Einstein - non pas l'Einstein historique, mais celui de notre expérience de pensée - regarde une fenêtre de conversation encore vide.

Il pourrait commencer par une question de physique. Il pourrait aussi soumettre une page inachevée, demander une objection, comparer deux formulations. Rien ne nous autorise à décider ce qu'il écrirait réellement. Le détail important est ailleurs : chaque fois qu'il utiliserait l'assistant, il laisserait une nouvelle trace. Une question datée. Une équation copiée. Une réponse acceptée, ignorée ou corrigée. Peut-être une série de tentatives qui montrerait moins une solution qu'une manière de chercher.

Supposons maintenant que cet assistant ne reçoive pas une seule question, mais un ensemble organisé de documents : articles, brouillons, carnets, lettres, conférences et corrections. Les archives éditées d'Einstein attestent l'existence d'un héritage écrit considérable, composé de textes scientifiques, de correspondances et d'autres documents situés dans le temps [\[SRC-001\]](#) [\[SRC-032\]](#) [\[SRC-036\]](#). Notre scénario ajoute à ces archives historiques

des interactions contemporaines entièrement imaginées.

L'assistant pourrait retrouver un passage, rapprocher deux formulations ou signaler qu'une idée apparaît sous plusieurs formes. Il pourrait apprendre des régularités de vocabulaire, de structure et d'argumentation. Si on lui présentait suffisamment d'exemples, il pourrait peut-être produire un texte qui ressemble, sous certains critères, aux textes observés.

Mais qu'aurait-il appris exactement ?

La tentation du raccourci

Nous sommes prompts à répondre : « Il aurait appris Einstein. » La formule est séduisante parce qu'elle contracte toute une chaîne d'opérations en trois mots. Elle est aussi dangereuse.

Le système n'aurait pas accès à une pensée avant qu'elle soit exprimée. Il recevrait des inscriptions : des caractères, des images numérisées, des métadonnées, peut-être des transcriptions. Il ne verrait pas l'instant où une intuition apparaît, la sensation d'une difficulté, l'attention qui se déplace ou la raison pour laquelle une phrase reste dans un tiroir. Même un corpus immense ne contient que ce qui a été formulé, conservé, retrouvé, sélectionné et rendu lisible.

Dire cela ne rend pas les traces insignifiantes. Une rature peut montrer qu'une formulation a été rejetée. Une suite de brouillons peut documenter une transformation. Une lettre peut replacer un problème dans une conversation. Ces éléments soutiennent des inférences. Ils ne suppriment pas la distance entre le document et l'activité mentale dont il est issu.

Cette distance constitue le premier fil du livre :

une IA ne lit pas directement un esprit ; elle traite des traces accessibles et peut construire des représentations utiles pour prédire, classer, inférer, générer ou simuler certains aspects de ce qui a été observé.

Le verbe *peut* est essentiel. La portée d'une représentation dépend des données, de l'architecture, de l'objectif d'apprentissage, des outils, du contexte et de la manière dont on l'évalue. Une imitation convaincante dans une situation ne garantit ni la fidélité générale du modèle ni une identité avec la personne.

Quatre mécanismes derrière un même écran

Dans notre scène fictive, Einstein écrit dans une seule fenêtre. Cette unité visuelle peut masquer plusieurs mécanismes.

Le modèle possède d'abord des **paramètres**, ajustés au cours d'un entraînement antérieur. La conversation en cours fournit ensuite un **contexte**, utilisé pour calculer la réponse présente. Un produit peut aussi conserver une **mémoire persistante** et la réintroduire dans des échanges ultérieurs. Enfin, un système peut effectuer une **recherche externe** dans des documents et placer les passages retrouvés dans le contexte.

Ces quatre éléments ne sont pas interchangeables. Une phrase saisie dans une conversation ne réécrit pas immédiatement, par sa seule présence, les paramètres du modèle général. Dans l'évaluation few-shot décrite pour GPT-3, les tâches sont réalisées sans mise à jour de gradient [SRC-009]. Chez OpenAI, la documentation d'ajustement supervisé décrit une opération différente : un jeu d'exemples est téléversé, puis un travail d'entraînement met à jour les poids du modèle [SRC-039]. Les politiques d'un service peuvent prévoir qu'une partie de certaines données soit utilisée plus tard dans un processus d'amélioration ; cela relève d'une autre temporalité, de contrôles et de choix de traitement. La documentation actuelle de ChatGPT distingue notamment les contrôles de mémoire des contrôles concernant l'utilisation de conversations pour améliorer les modèles [SRC-006] [SRC-007]. Cette description est datée : elle caractérise un produit et une politique susceptibles d'évoluer, non une loi universelle de l'intelligence artificielle.

Notre Einstein imaginé pourrait donc avoir sous les yeux un assistant dont les réponses combinent plusieurs sources d'information. Pour savoir ce que le système a réellement utilisé, l'apparence de la réponse ne suffirait pas. Il faudrait documenter la configuration.

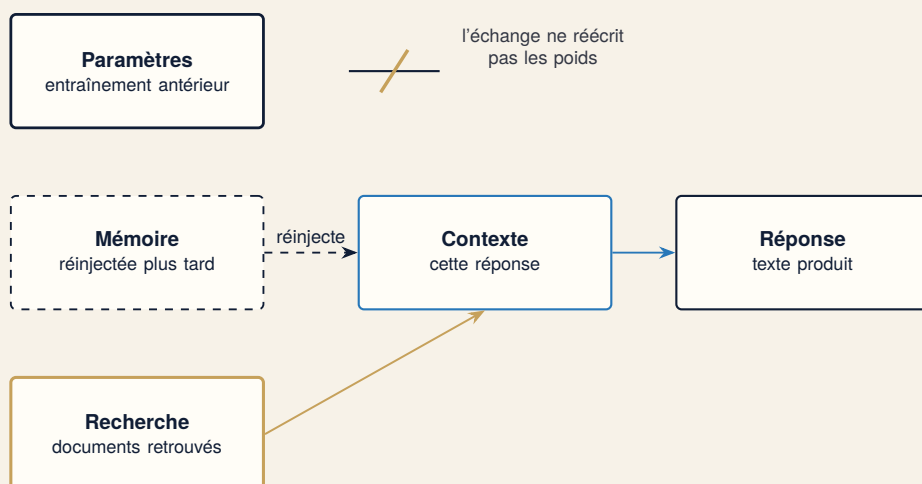


Figure. Paramètres, contexte, mémoire de produit et recherche externe n'entrent pas dans le même geste. L'échange alimente le contexte. Il ne met pas à jour les poids. Schéma explicatif, non un résultat expérimental. [SRC-006] [SRC-007] [SRC-009] [SRC-039], aux niveaux d'examen du registre, documentation produit datée du 9 octobre 2026.

■ ÉTABLI

Les paramètres appris, le contexte courant, la mémoire persistante d'un produit et la recherche de documents externes sont des mécanismes distincts. Une conversation ne doit pas être décrite comme une modification instantanée du modèle général.

Le double test

Imaginons qu'après des mois d'interaction, l'assistant reçoive un problème nouveau. Il propose une réponse. Des spécialistes la jugent cohérente avec certaines habitudes observées dans les écrits d'Einstein. Que prouverait ce succès ?

Il montrerait d'abord une performance : dans les conditions du test, le système a produit quelque chose qui satisfait des critères définis. Si les documents utilisés pour l'évaluation avaient été tenus à l'écart de la construction du modèle, le résultat serait plus informatif que la simple répétition d'un passage connu. Si plusieurs évaluateurs, plusieurs problèmes et plusieurs formulations convergeaient, l'inférence gagnerait en solidité.

Elle resterait néanmoins limitée. Un texte peut ressembler à un autre pour de mauvaises raisons. Un modèle peut exploiter des indices superficiels. Les évaluateurs peuvent partager les mêmes attentes. Surtout, la réussite ne démontrerait pas que le système éprouve le problème comme Einstein l'aurait éprouvé, ni qu'il possède son identité, ses souvenirs ou sa conscience.

Le test devrait donc être double. Il faudrait mesurer ce que le modèle reproduit et rechercher activement où il échoue. Quels changements de formulation brisent sa cohérence ? Quelles affirmations contredisent les documents ? Quels aspects de la démarche scientifique sont absents ? Quand produit-il une réponse plausible mais sans provenance ?

Une représentation sérieuse ne se juge pas seulement à ses ressemblances. Elle se juge aussi à la carte de ses limites.

Deux miroirs

L'expérience de pensée commence avec un individu, mais elle ne peut pas y rester. Les systèmes contemporains sont entraînés à partir

d'ensembles qui rassemblent des productions issues de nombreuses personnes, institutions, langues, disciplines et périodes. Des textes, des images, du code et des mesures deviennent des données après des opérations de collecte, de filtrage et de transformation.

Le premier miroir de ce livre suit la chaîne :

individu → pensée exprimée → traces → données → modèle de certains aspects de la personne.

Le second l'élargit :

humanité → observations et productions → données → apprentissage → représentations utilisées pour prédire et générer.

Dans les deux cas, un même avertissement s'impose : la carte n'est pas le territoire. Les données ne sont pas la totalité de ce qu'elles représentent. Elles peuvent être abondantes et néanmoins lacunaires, précises et néanmoins sélectionnées, utiles et néanmoins biaisées.

La comparaison entre les deux miroirs ne suppose pas qu'une personne et le monde soient modélisés de la même façon. Elle sert à poser une question commune : que conserve une représentation lorsqu'elle est construite à partir de traces, et que perd-elle ?

Ce que la scène nous oblige à demander

Revenons à la fenêtre blanche. L'Einstein fictif de 2026 pourrait fermer l'ordinateur, reprendre son carnet et laisser la réponse sans commentaire. Ce silence serait lui aussi ambigu. Signifierait-il l'accord, le doute, la fatigue ou la décision de chercher autrement ? Sans trace supplémentaire, nous ne le saurions pas.

Cette ignorance n'est pas un défaut provisoire que davantage de calcul efface nécessairement. Elle rappelle que toute connaissance d'une personne passe par des observations et des interprétations. Les humains vivent déjà avec cette limite lorsqu'ils lisent une lettre, reconstituent une découverte ou tentent de comprendre un proche. L'intelligence artificielle ne crée pas ce problème. Elle l'agrandit, l'automatise et nous force à préciser nos critères.

Les chapitres qui suivent commenceront donc avant ChatGPT. Ils demanderont ce que nous appelons une pensée, comment elle laisse des traces et ce que signifie connaître quelqu'un à partir de ses productions. Ce détour est indispensable. Sans lui, nous risquerions de confondre

la fluidité d'une réponse avec un accès à l'esprit qui l'a inspirée. Cette prudence sur le vocabulaire employé pour parler des modèles de langage est elle-même défendue dans la littérature contemporaine [SRC-019] [SRC-025].

À RETENIR

- 01 Le prologue est une expérience de pensée, non un récit historique.
- 02 Un assistant traite des traces accessibles, pas un esprit directement.
- 03 Paramètres, contexte, mémoire et recherche externe doivent rester distincts.
- 04 Une imitation se valide par des tests et par l'étude de ses échecs.
- 05 La question individuelle ouvre sur celle des représentations du monde construites à partir de traces humaines.

Sources mobilisées

- **SRC-001 - Einstein Papers Project.** Portée institutionnelle des écrits, correspondances et documents édités. Consultation de la présentation officielle et du portail documentaire.
- **SRC-006 - OpenAI, Data Controls FAQ.** Documentation produit consultée le 9 octobre 2026 ; politique datée et susceptible d'évoluer.
- **SRC-007 - OpenAI, Memory FAQ.** Documentation produit consultée le 9 octobre 2026 ; description de la mémoire comme mécanisme de personnalisation.
- **SRC-009 - Brown et al., apprentissage few-shot.** Notice et résumé examinés ; utilisé pour la réalisation de tâches sans mise à jour de gradient dans l'évaluation décrite.
- **SRC-039 - OpenAI, ajustement supervisé.** Page officielle consultée le 9 octobre 2026 ; elle décrit la mise à jour des poids comme le résultat d'un travail d'entraînement, non d'un message isolé. La disponibilité de cette plateforme évolue.
- **SRC-019 - Mitchell et Krakauer, débat sur la compréhension des modèles de langage.** Notice et résumé vérifiés ; article de perspective, non verdict expérimental unique.
- **SRC-025 - Shanahan, langage employé à propos des grands modèles de langage.** Métadonnées, résumé et extraits de conclusion examinés.

- **SRC-032 - Einstein Papers Project, histoire du projet.** Page institutionnelle intégralement consultée; elle décrit l'organisation et l'édition des papiers après 1955.
- **SRC-036 - Collected Papers, volume 5.** Notice éditoriale officielle consultée; elle décrit la portée de plus de 500 lettres sans autoriser de citation individuelle.
- **SRC-038 - Brown et Fehige, expériences de pensée.** Entrée académique intégralement consultée; elle encadre le statut méthodologique du scénario imaginé.

1 · Qu'est-ce qu'une pensée ?

Avant de demander à une machine de modéliser une pensée, il faut préciser ce que ce mot nous autorise réellement à observer.

Un mot pour plusieurs phénomènes

« J'ai une pensée. » La phrase semble désigner quelque chose de simple. Elle peut pourtant renvoyer à une image mentale, une phrase intérieure, un souvenir, une intention, une comparaison, un calcul, une émotion interprétée ou une question encore sans mots. Le même terme rassemble des phénomènes différents.

Cette diversité crée une difficulté immédiate. Si nous cherchons une définition assez large pour tout inclure, elle risque de ne plus rien expliquer. Si nous choisissons une définition trop étroite - par exemple « une pensée est une phrase intérieure » - nous excluons les raisonnements visuels, les gestes préparés sans verbalisation ou les intuitions qui précèdent leur formulation.

Le but de ce chapitre n'est donc pas de résoudre le problème philosophique de la pensée. Il est plus modeste et plus utile pour notre enquête : construire des distinctions qui empêchent de confondre un phénomène mental, son expression et le modèle calculé à partir de cette expression.

Nous utiliserons le mot **pensée** pour désigner une activité ou un état mental attribué à une personne. Cette définition reste volontairement générale. Elle ne tranche ni la nature de la conscience, ni la manière dont une expérience subjective émerge, ni la frontière exacte entre perception, émotion et raisonnement. Ces questions sont importantes ; elles dépassent ce que les preuves mobilisées ici permettent d'établir.

■ INCONNU

Aucune définition brève de la pensée ne résout à elle seule le problème de la conscience. Le livre adopte une définition opérationnelle pour étudier les relations entre activités mentales, expressions observables et modèles.

Trois plans à ne pas confondre

Prenons une situation ordinaire. Une personne hésite devant deux itinéraires, regarde le ciel, puis dit : « Je vais prendre le train. » Nous pouvons distinguer au moins trois plans.

Le premier est celui de l'**expérience et de l'activité mentales**. La personne peut se souvenir d'un retard, imaginer la pluie, estimer une durée ou ressentir une préférence. Une partie de cette activité lui est peut-être accessible ; une autre ne l'est pas clairement.

Le deuxième est celui de l'**expression observable**. La phrase prononcée, le regard vers le ciel, le billet acheté ou le trajet choisi sont accessibles à autrui. Ils permettent des inférences : la pluie a peut-être compté, le temps de parcours aussi. Mais plusieurs histoires différentes peuvent produire le même choix.

Le troisième est celui de la **description ou du modèle**. Un observateur peut résumer : « cette personne évite le risque météorologique ». Un système peut estimer une probabilité de choisir le train. Dans les deux cas, une représentation est construite à partir d'indices. Elle peut être utile sans être exhaustive.

Ces plans se répondent, mais ils ne sont pas identiques. Une expression est produite par une personne dans une situation ; elle n'est pas une copie complète de ce qui se passe en elle. Un modèle organise des observations ; il n'est pas la personne observée.



Activité mentale traits incomplets : non observée directement



Expression trait plein : parole, geste, écriture



Modèle filet or : description construite

Figure. Trois plans reliés et distincts : l'activité mentale n'est pas observée directement, l'expression peut l'être, le modèle est construit ensuite. Schéma de l'auteur.

Un exemple suivi : corriger une équation

Imaginons une page de travail. Une équation y figure. Une ligne est barrée. Une autre formulation apparaît à côté. Aucune lettre n'explique le geste. Trois lectures restent possibles, et elles n'ont pas le même statut.

Observation. Une équation a été écrite, puis corrigée sur le support.

Inférence. L'auteur a modifié sa formulation. Plusieurs causes restent compatibles avec cette seule page : erreur détectée, simplification, anticipation d'une objection, ou simple manque de place.

Hypothèse. Cette correction pourrait témoigner d'une évolution dans sa manière de traiter le problème. Cette hypothèse n'est pas fausse par principe. Elle n'est pas non plus établie par la rature seule. Il faudrait des pièces environnantes : une version datée, une lettre, une publication ultérieure.

Inconnu. La raison exacte de la correction, l'état subjectif au moment du geste, et ce qui n'a jamais été écrit restent hors de portée des pièces disponibles.

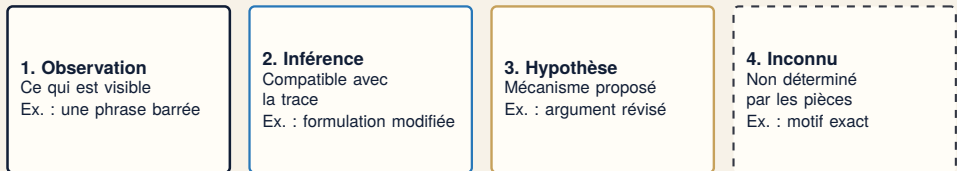


Figure. Quatre niveaux pour lire une trace : observation, inférence, hypothèse, inconnu. Ce principe méthodologique revient dans tout le livre. Schéma de l'auteur.

Cette grille - observation, inférence, hypothèse, inconnu - accompagnera le reste du livre. Elle s'appliquera aux archives d'Einstein, aux réponses d'un assistant et aux affirmations que nous serons tentés de lire comme des preuves.

La question centrale n'est donc pas : « qu'est-ce que la pensée en général ? » Elle devient opérationnelle : **qu'est-ce qu'une IA peut raisonnablement apprendre des traces laissées par une personne, sans confondre ces traces avec la personne elle-même ?**

Pensée exprimée, pensée attribuée

Nous n'attendons pas une technologie avancée pour attribuer des pensées. Nous le faisons continuellement. Une amie cherche ses clés : nous supposons qu'elle veut sortir. Un collègue relit trois fois une phrase : nous inférons qu'il doute de sa formulation. Ces attributions permettent la coordination sociale.

Elles peuvent aussi être fausses. La personne cherchait peut-être un document posé près des clés. Le collègue corrigeait une faute sans remettre l'idée en question. Une observation sous-détermine souvent l'explication : plusieurs états possibles sont compatibles avec la même trace.

Cette sous-détermination ne condamne pas toute tentative de connaissance d'autrui. Elle impose toutefois une méthode. Une attribution gagne en crédibilité lorsqu'elle s'appuie sur des indices indépendants, résiste à des explications concurrentes, prédit des observations nouvelles et peut être révisée. Elle reste néanmoins une inférence, pas un accès sans médiation. Les recherches sur la mentalisation étudient précisément cette capacité à représenter ou inférer les états mentaux d'autrui à partir d'indices et d'un point de vue [SRC-022]. Cette idée possède une histoire plus ancienne dans les travaux sur la « théorie de l'esprit », définie comme l'attribution d'états mentaux non directement observables afin d'expliquer ou de prédire un comportement [SRC-027].

Le langage ajoute un cas particulier : une personne peut décrire ce qu'elle pense. Ce témoignage est précieux. Il peut préciser une intention invisible dans le comportement. Mais il est lui aussi produit dans une situation, avec un vocabulaire, un destinataire et un but. Nous pouvons oublier, simplifier, réorganiser ou découvrir après coup une explication qui nous paraît cohérente. Prendre au sérieux un récit subjectif ne signifie pas le traiter comme une transcription parfaite et complète.

Représenter n'est pas photographier

Le terme **représentation mentale** peut suggérer une petite image rangée dans la tête. Cette métaphore convient parfois à l'intuition, mais elle devient vite trompeuse. Une représentation peut être une structure qui permet de reconnaître une situation, de comparer des possibilités ou de guider une action sans ressembler visuellement à ce qu'elle représente.

Les sciences cognitives emploient plusieurs cadres pour étudier ces structures. La présentation éditoriale de l'ouvrage de Johnson-Laird décrit une théorie dans laquelle une personne construirait des répliques intérieures des relations entre objets et événements, puis raisonnerait à partir de ces modèles. [SRC-018] Cette formulation vient de la notice de l'éditeur : le livre n'a pas été lu intégralement pour cet audit. D'autres travaux distinguent plusieurs systèmes de mémoire ou étudient la construction des souvenirs autobiographiques. [SRC-016] [SRC-017] Ces approches ne se réduisent pas à une théorie unique de la pensée. Elles fournissent des concepts testables pour des phénomènes particuliers.

Le mot **représentation** doit donc toujours appeler deux questions : utile à quoi, et mise en évidence comment ? Une structure interne peut être qualifiée de représentation si elle varie de manière systématique avec certaines informations et contribue à un comportement ou une prédiction. Ce

constat ne dit pas encore si elle est consciente, vécue ou comprise comme elle le serait par un humain.

Cette prudence sera décisive lorsque nous parlerons des réseaux de neurones. Le fait qu'un modèle encode des régularités ne suffit pas à conclure qu'il pense. Inversement, refuser tout usage du mot « représentation » parce que le système n'est pas humain empêcherait d'étudier ce que ses états internes permettent effectivement de faire.

La mémoire n'est pas une archive intérieure

Nous imaginons volontiers un souvenir comme un document conservé puis rouvert. L'analogie est pratique : elle donne un début, un contenu et un lieu. Elle peut cependant masquer le caractère reconstitutif du rappel autobiographique. Les travaux sur la mémoire du soi décrivent le souvenir comme le résultat d'une interaction entre connaissances autobiographiques, buts présents et indices de récupération, plutôt que comme la lecture inchangée d'un enregistrement. [SRC-017] La psychologie de la mémoire documente aussi plusieurs formes d'oubli, de déformation et d'attribution erronée. Ces limites ne rendent pas la mémoire inutile, mais interdisent de la traiter comme une copie [SRC-021].

Cette idée a deux conséquences pour notre projet. Premièrement, les récits qu'une personne produit sur elle-même sont des données importantes, mais situées. Ils peuvent varier sans que l'une des versions soit nécessairement un mensonge. Deuxièmement, un modèle entraîné sur ces récits apprend les régularités des récits disponibles. Il n'obtient pas pour autant un accès à un événement mental original qui existerait intact derrière eux. Une IA peut donc apprendre des formes de mise en récit sans apprendre l'expérience qui les a rendues nécessaires.

Ces récits ne sont pas seulement des rapports sur le passé. Les recherches sur l'identité narrative les décrivent aussi comme une manière évolutive d'articuler un passé reconstruit et un futur imaginé [SRC-026]. Un modèle qui apprend ces récits peut donc capter des formes de mise en cohérence sans posséder l'histoire vécue qui leur donne sens.

La différence entre mémoire humaine et mémoire informatique doit également rester visible. Dans un produit numérique, « mémoire » peut désigner des informations persistantes enregistrées puis réinjectées dans un contexte. Chez une personne, le mot renvoie à des processus biologiques et cognitifs multiples. Employer le même terme ne prouve aucune identité de mécanisme.

Ce que la conscience n'a pas à résoudre ici

Ce livre n'a pas besoin de trancher les débats sur la conscience pour avancer. Une règle suffit : ne pas utiliser une performance observable comme preuve automatique d'une expérience subjective. Un système peut classer, prédire ou générer. Ces capacités sont mesurables. Savoir si elles s'accompagnent d'une expérience comparable à la nôtre est une question distincte, et non résolue pour les modèles étudiés ici.

■ ÉTABLI

Une performance comportementale et une attribution d'expérience subjective sont deux propositions distinctes. La première peut être mesurée par un protocole ; la seconde exige des critères supplémentaires et reste controversée lorsqu'elle concerne des systèmes artificiels.

Une définition de travail

Nous pouvons maintenant formuler la définition qui guidera les chapitres suivants :

Dans ce livre, une pensée désigne une activité ou un état mental attribué à une personne. Nous n'en observons directement que certaines expressions et certains effets. À partir d'eux, nous construisons des descriptions et des modèles révisables.

Cette formulation ne nie pas l'accès qu'une personne possède à sa propre expérience. Elle précise le point de vue de l'enquête : celui d'un observateur, d'un historien, d'un proche ou d'un système qui travaille à partir de traces.

Elle évite aussi deux excès symétriques. Le premier consisterait à dire que, puisque nous n'accédons pas directement à l'esprit d'autrui, aucune connaissance n'est possible. Nos relations, la psychologie et l'histoire montrent au contraire que des inférences peuvent être robustes. Le second consisterait à croire qu'un grand nombre de traces abolit toute incertitude. La quantité améliore parfois l'estimation ; elle ne transforme pas les observations en identité.

Le chapitre suivant s'intéresse au passage décisif entre le premier et le deuxième plan : comment une activité humaine devient-elle une trace que d'autres pourront conserver, classer et interpréter ?

À RETENIR

-
- 01 Le livre adopte une définition opérationnelle de la pensée, utile à l'enquête.
 - 02 Observation, inférence, hypothèse et inconnu ne se confondent pas.
 - 03 Activité mentale, expression et modèle sont trois plans distincts.
 - 04 Une IA peut apprendre des régularités de traces sans accéder à la personne.
 - 05 La conscience n'est pas tranchée par une performance observable.
-

Sources mobilisées

- **SRC-016 - Tulving, mémoire épisodique et sémantique.** Source à utiliser seulement selon le niveau d'examen enregistré dans le registre.
- **SRC-017 - Conway et Pleydell-Pearce, système mémoire-soi.** Sert à encadrer le caractère construit du souvenir autobiographique.
- **SRC-018 - Johnson-Laird, modèles mentaux.** Présentation éditoriale consultée; l'ouvrage intégral n'a pas été lu. La phrase du chapitre reprend cette présentation, non une démonstration vérifiée page à page.
- **SRC-019 - Mitchell et Krakauer, débat sur la compréhension des modèles de langage.** Notice et résumé vérifiés; utilisé pour distinguer capacités et formes possibles de compréhension.
- **SRC-021 - Schacter, erreurs de mémoire.** Notice et résumé examinés; utilisé pour encadrer le caractère faillible et constructif du rappel.
- **SRC-022 - Frith et Frith, mentalisation.** Métadonnées, résumé et extraits examinés; utilisé pour définir l'inférence sur les états mentaux d'autrui.
- **SRC-026 - McAdams et McLean, identité narrative.** Notice DOI et résumé éditeur examinés; utilisé pour relier passé reconstruit, futur imaginé et récit de soi.
- **SRC-027 - Premack et Woodruff, théorie de l'esprit.** Notice et résumé Cambridge examinés; étude historique sur un chimpanzé, non lecture directe de l'esprit humain.

2 · L'esprit laisse des traces

Une rature, une lettre ou une équation peuvent documenter un chemin intellectuel - à condition de ne pas les prendre pour le chemin lui-même.

Ce qui arrive jusqu'à nous

Une pensée n'arrive pas dans les archives. Ce qui y parvient, c'est une phrase recopiée, une équation barrée, une lettre envoyée, parfois le récit d'un témoin. Entre l'activité mentale et le document, plusieurs transformations ont déjà eu lieu : choisir une expression, l'inscrire sur un support, conserver ce support, le classer, puis le rendre accessible.

Ces traces comptent. Une correction révèle qu'une formulation a été abandonnée. Deux versions permettent de suivre une évolution. Une correspondance replace une décision dans un échange. Mais aucune pièce ne contient, à elle seule, la totalité de l'expérience qui l'a produite. Le silence, l'hésitation ou une intuition jamais formulée peuvent ne laisser aucune marque.

Les archives d'Einstein rendent cette différence concrète. L'Einstein Papers Project édite des écrits, des lettres et des documents liés à sa vie et à son travail [\[SRC-001\]](#) [\[SRC-032\]](#). La notice officielle du volume 5 décrit à elle seule plus de cinq cents lettres pour les années 1902 à 1914 [\[SRC-036\]](#). Celle du volume 4 signale des notes et des calculs qui documentent une progression de recherche vers une théorie relativiste de la gravitation [\[SRC-035\]](#). Cette abondance permet d'étudier des chronologies, des formulations et des échanges. Elle ne transforme pas l'archive en enregistrement continu de la pensée d'Einstein.

Ce chapitre répond à une question précise, distincte de celle du suivant : **que reste-t-il d'une activité humaine dans les traces conservées ?** La transformation documentaire est ici l'objet principal. L'inférence sur l'auteur viendra au chapitre 3.

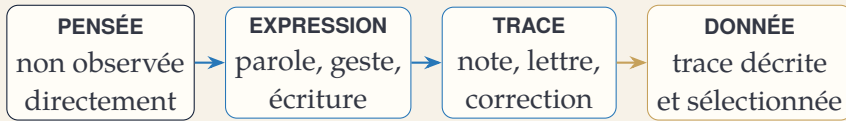


Figure. Une pensée exprimée peut laisser une trace ; une trace décrite et sélectionnée peut devenir une donnée. Chaque passage transforme ce qui sera ensuite interprété. Schéma conceptuel de l'auteur.



Chaque flèche facilite l'étude et peut aussi perdre du contexte, de la nuance ou de la provenance.

Figure. De l'expression au corpus exploitable. Chaque transformation facilite l'étude et peut modifier ou perdre de l'information. Schéma conceptuel de l'auteur.

Exprimer transforme déjà

Passer d'une idée à une phrase n'est pas une opération neutre. Il faut choisir un vocabulaire, un ordre, un niveau de détail et un destinataire. Une note pour soi ne ressemble pas à un article. Une lettre à un collègue n'a pas la même fonction qu'une conférence publique. La même personne peut présenter le même problème de plusieurs manières sans que l'une d'elles soit la version définitive de sa pensée.

Cette transformation est particulièrement visible dans les brouillons. Une rature montre un changement sur la page, mais elle n'en donne pas automatiquement la raison. L'auteur a-t-il découvert une erreur, préféré une formulation plus claire, anticipé une objection ou simplement manqué de place ? Le document réduit l'ensemble des explications possibles ; il ne choisit pas toujours entre elles.

Le contexte peut aider. Une lettre écrite le même jour, une version suivante ou la réponse d'un correspondant ajoutent des indices. L'interprétation devient alors un travail de mise en relation. Elle ne consiste pas à faire parler une trace isolée, mais à comparer des pièces dont la provenance et la date sont connues.

Cette règle vaut aussi pour les traces numériques. Un message très court peut être une demande urgente, un rappel ou la suite d'une conversation orale. Un historique de modifications montre ce qui a changé, pas nécessairement pourquoi. Plus une trace est facile à collecter, plus nous devons résister à l'idée qu'elle est auto-explicative.

Conserver, c'est sélectionner

Nous parlons des archives comme si elles recevaient passivement le passé. En réalité, toute conservation dépend de décisions et de circonstances. Certains documents sont jetés, perdus ou détruits. D'autres sont conservés parce qu'ils semblent importants, parce qu'une institution les acquiert ou simplement parce que leur support résiste.

La sélection continue après la conservation. Il faut inventorier, décrire, numériser et parfois traduire. Une édition scientifique choisit un ordre, établit un texte, relie des variantes et ajoute des annotations. Ce travail augmente considérablement l'utilité du corpus. Il ajoute aussi une couche éditoriale qui doit rester visible.

Dans le cas d'Einstein, les *Collected Papers* ne sont pas un amas de fichiers anonymes. Ils constituent une édition savante produite par un projet institutionnel, avec des volumes, des commentaires et des liens vers des documents d'archives [\[SRC-001\]](#) [\[SRC-033\]](#). La différence est fondamentale pour un système d'intelligence artificielle : un passage accompagné de sa date, de son auteur, de sa langue et de son statut documentaire est plus interprétable qu'une chaîne de caractères détachée de sa provenance.

■ ÉTABLI

Une archive éditée est un corpus construit : la conservation, la description, l'ordre et l'annotation résultent de choix documentaires. Ces choix améliorent l'accès tout en délimitant ce qui peut être observé.

Quand une trace devient une donnée

Une trace n'est pas automatiquement une donnée prête pour l'apprentissage. Pour être traitée par un système, elle doit être représentée sous une forme exploitable. Une page peut être photographiée, son texte transcrit, sa langue identifiée, sa date encodée et ses relations avec d'autres documents décrites.

Chaque opération ouvre des possibilités et introduit des erreurs. Une page manuscrite peut passer par la conservation physique, la numérisation, la reconnaissance optique, la transcription, la normalisation, l'indexation, puis l'intégration dans un corpus. Un système qui ignore ces incertitudes reçoit une version artificiellement nette du matériau.

Étape	Gain	Risque
Numérisation	Accès et reproduction	Perte de détails matériels
Reconnaissance optique	Texte interrogeable	Confusion de caractères
Transcription	Texte exploitable	Normalisation ou omission
Traduction	Accessibilité linguistique	Perte de nuances
Métadonnées	Contexte et classement	Attribution ou datation erronée
Indexation	Recherche rapide	Relations mal représentées

Tableau. Grille d'analyse des transformations documentaires. Chaque étape peut améliorer l'accès tout en introduisant une ambiguïté.

La qualité ne se résume donc pas au nombre de documents. Elle dépend de la provenance, de la granularité, des métadonnées, des droits d'utilisation et de la capacité à retrouver le support original. Un million de fragments sans contexte peuvent être moins utiles pour certaines questions qu'un ensemble plus petit, daté et annoté. Ce que l'assistant du chapitre 5 pourra « apprendre » dépendra directement de ces transformations.

Il faut également distinguer la donnée d'entraînement du document consulté au moment d'une réponse. Dans le premier cas, des exemples participent à une procédure qui ajuste des paramètres. Dans le second, un moteur retrouve un document et le place dans le contexte. Le texte source peut être identique ; son rôle technique ne l'est pas.

Les traces négatives

Une archive montre ce qui est présent. Pour raisonner correctement, il faut aussi considérer ce qui manque. L'absence d'une lettre peut signifier qu'elle n'a jamais été écrite, qu'elle a été perdue, qu'elle n'a pas encore été retrouvée ou qu'elle n'est pas accessible. Ces hypothèses n'ont pas les mêmes conséquences. Un témoignage rétrospectif ne comble pas automatiquement ce vide : le rappel humain est lui aussi exposé à l'oubli et à l'attribution erronée [\[SRC-021\]](#).

La même prudence vaut pour les régularités. Si une personne n'aborde jamais un sujet dans le corpus, nous ne pouvons pas conclure qu'elle n'y a jamais pensé. Si elle emploie toujours un mot dans des textes publics, nous ne savons pas si elle l'utilise de la même manière dans ses notes privées. Un corpus décrit une distribution de traces, pas l'ensemble de toutes les situations possibles.

Ces absences peuvent être systématiques. Les activités considérées comme ordinaires laissent parfois moins de documents que les événements publics. Les personnes disposant de moyens institutionnels sont mieux archivées. Certaines langues, certains supports et certains groupes survivent davantage dans les collections. Un modèle construit sur des

archives hérite de cette géographie inégale.

Reconnaître ces limites ne condamne pas le projet. Cela permet de poser des questions plus précises : quelles périodes sont couvertes ? Quels genres de documents dominent ? Quels intermédiaires ont sélectionné les pièces ? Quels silences sont connus, et lesquels restent invisibles ?

Lire un chemin sans inventer ses étapes

Supposons que trois versions d'un raisonnement soient conservées. La première contient une analogie, la deuxième une équation supplémentaire, la troisième supprime l'analogie. Nous pouvons établir une séquence documentaire si les dates ou les indices matériels le permettent. Nous pouvons décrire les différences. Nous pouvons parfois relier un changement à une lettre ou à une expérience connue.

Nous devons être plus prudents lorsque nous racontons ce qui s'est passé « dans la tête » de l'auteur. Dire qu'il a supprimé l'analogie est une observation. Dire qu'il la jugeait trompeuse est une inférence. Dire qu'une intuition précise lui est apparue à cet instant peut être une hypothèse sans preuve.

Le vocabulaire épistémique du livre sert exactement à cela. Il n'interdit pas le récit ; il indique sa solidité. Une histoire des idées devient plus intéressante, non moins, lorsque le lecteur sait où finit le document et où commence l'interprétation.

Ce qu'un assistant pourrait exploiter

Un assistant relié à un corpus bien documenté pourrait accomplir plusieurs tâches utiles. Il pourrait retrouver toutes les occurrences d'un terme, rapprocher des versions, construire une chronologie provisoire ou signaler une contradiction apparente. Il pourrait proposer des passages pertinents à un historien, qui retournerait aux documents.

Il pourrait aussi générer une synthèse erronée, fusionner deux périodes ou présenter une inférence comme un fait. La fluidité du texte ne garantit pas la fidélité archivistique. Pour cette raison, chaque réponse importante devrait conserver un chemin vers les pièces utilisées et distinguer citation, paraphrase et interprétation.

Un modèle personnalisé pourrait apprendre des régularités plus abstraites : longueur des phrases, ordre des arguments, types d'exemples ou manière de répondre à une objection. Ces régularités appartiennent aux productions observées. Les appeler « style de pensée » peut être utile dans un sens limité, à condition de ne pas les confondre avec la totalité de la

pensée.

■ INFÉRENCE

Des traces nombreuses, datées et diversifiées peuvent soutenir un modèle plus robuste de certaines régularités intellectuelles. Elles ne garantissent ni l'exhaustivité du corpus ni l'accès à l'expérience subjective qui a accompagné leur production.

De la trace au modèle

Le passage décisif peut maintenant être formulé. Une pensée exprimée produit parfois une trace. Cette trace peut être conservée et décrite. Une sélection de traces peut devenir un corpus. Un système peut enfin construire, à partir de ce corpus, une représentation qui prédit ou génère certaines productions.

À chaque étape, quelque chose est gagné : stabilité, comparabilité, recherche, calcul. À chaque étape, quelque chose peut être perdu : contexte, ambiguïté, silence, provenance ou singularité. La rigueur consiste à documenter ces transformations plutôt qu'à les effacer sous le mot « données ».

Le prochain chapitre partira du corpus ainsi constitué. Même si les traces sont authentiques et soigneusement éditées, jusqu'où permettent-elles de connaître une personne ? Comment distinguer un modèle utile d'un portrait circulaire qui ne retrouve que les attentes de ses concepteurs ?

À RETENIR

- 01 Ce chapitre étudie la transformation documentaire, pas encore l'inférence sur l'auteur.
- 02 Chaque étape d'accès peut aussi perdre, modifier ou rendre ambigu.
- 03 La provenance et les métadonnées comptent autant que le volume.
- 04 L'absence d'une trace n'est pas la preuve de l'absence d'une pensée.
- 05 Ce qu'un système pourra exploiter dépend de ces transformations.

Sources mobilisées

- **SRC-001 - Einstein Papers Project et Einstein Database.** Présentations institutionnelles consultées ; elles étaient la portée du corpus, son orga-

nisation savante et les catégories générales de documents.

- **SRC-021 - Schacter, erreurs de mémoire.** Notice et résumé examinés ; utilisé seulement pour rappeler les limites du témoignage rétrospectif.
- **SRC-004 - Lewis et al., génération augmentée par recherche.** Résumé vérifié ; utilisé uniquement pour distinguer récupération documentaire et mémoire paramétrique.
- **SRC-032 - Einstein Papers Project, histoire du projet.** Page institutionnelle intégralement consultée ; elle documente l'histoire de l'organisation et de l'édition des papiers.
- **SRC-033 - Collected Papers, volume 1.** Notice éditoriale officielle consultée ; elle caractérise la portée de la correspondance de jeunesse sans attribuer de contenu non vérifié.
- **SRC-035 - Collected Papers, volume 4.** Notice éditoriale officielle consultée ; elle étaye la présence de notes et calculs de recherche pour 1912-1914.
- **SRC-036 - Collected Papers, volume 5.** Notice éditoriale officielle consultée ; elle étaye la portée générale de plus de 500 lettres, pas leur contenu individuel.
- **Sources archivistiques complémentaires.** À intégrer après vérification du lot de recherche de phase 3 ; aucune citation historique précise n'est attribuée sans document localisé.

3 · Peut-on connaître quelqu'un sans être lui ?

Un modèle peut devenir précis, utile et surprenant sans cesser d'être une inférence construite à partir d'un point de vue limité.

Après la trace, l'inférence

Le chapitre précédent a suivi le document : expression, conservation, numérisation, métadonnées, corpus. Celui-ci change d'objet. La question n'est plus : *que reste-t-il ?* Elle devient : **que pouvons-nous déduire de ces traces au sujet de leur auteur ?**

Nous connaissons des personnes que nous ne sommes pas. Nous anticipons la réaction d'un ami, reconnaissons le style d'une collègue et comprenons qu'une phrase ironique ne doit pas être prise au pied de la lettre. Cette connaissance n'est ni parfaite ni imaginaire. Elle s'appuie sur une histoire d'interactions, des paroles, des gestes et des situations partagées.

Elle reste pourtant indirecte. Nous n'éprouvons pas la douleur d'autrui à sa place. Nous ne voyons pas tous ses souvenirs au moment où elle décide. Nous construisons des hypothèses sur ses intentions et ses croyances, puis nous les révisons lorsque ses paroles ou ses actions les contredisent. Cette inférence sur les états mentaux d'autrui est étudiée sous le nom de mentalisation [SRC-022]. La stratégie consistant à expliquer ou prédire un comportement en attribuant croyances et désirs possède aussi une histoire philosophique [SRC-024].

La question n'est donc pas de choisir entre deux extrêmes - connaître parfaitement ou ne rien connaître. Elle est de déterminer ce qu'un modèle permet de prévoir, dans quelles situations, avec quelle incertitude et au prix de quelles simplifications.

Cette question précède l'intelligence artificielle. Un biographe, un médecin, un proche et un recruteur construisent tous des représentations partielles d'une personne. Leurs données, leurs buts et leurs responsabilités diffèrent. Un modèle calculé rend cette construction plus explicite et par-

fois plus facile à déployer à grande échelle. Il ne la rend pas automatiquement neutre.

Du portrait au modèle prédictif

Un portrait cherche souvent à rendre une personne intelligible. Il sélectionne des événements, des traits et des tensions. Un modèle prédictif poursuit un but plus opérationnel : à partir d'observations passées, estimer une réponse, un choix ou une préférence dans une situation définie.

La différence est importante. Un texte peut être un portrait convaincant sans permettre de prévoir une réaction nouvelle. Inversement, un modèle peut prédire correctement des choix limités sans raconter une histoire fidèle de la personne. Il peut exploiter des régularités statistiques dont l'interprétation humaine reste incertaine.

Des études sur les traces numériques montrent la puissance et l'étroitesse de ce type d'inférence. Dans des jeux de données définis, des « Likes » Facebook ont permis de prédire statistiquement certains attributs ou traits psychométriques [\[SRC-028\]](#) [\[SRC-029\]](#). Ces résultats portent sur des cibles, des populations et une plateforme précises. Ils ne démontrent ni une compréhension globale de la personne, ni la reconstruction de son histoire vécue.

Des travaux donnent une forme expérimentale à cette idée sans résoudre le cas humain. Rabinowitz et ses collègues ont entraîné un réseau à prédire des agents artificiels observés dans des mondes-grilles simples [\[SRC-023\]](#). Park et ses collègues ont construit un environnement où vingt-cinq agents génératifs combinaient un historique en langage naturel, une récupération de souvenirs, une étape de réflexion et une planification [\[SRC-012\]](#). Ces démonstrations portent sur des tâches et des architectures délimitées. Elles ne prouvent ni la reconstruction d'une personne réelle, ni l'accès à son expérience. Des modèles de langage ont aussi reproduit certaines distributions de réponses de groupes ou certains résultats expérimentaux [\[SRC-030\]](#) [\[SRC-031\]](#). Là encore, simuler une distribution ou une réponse moyenne ne suffit pas à établir la fidélité à un individu particulier.

Pour modéliser certaines productions intellectuelles, nous pourrions réunir des textes, des corrections, des réponses à des objections et des choix bibliographiques. Le système chercherait des régularités : concepts associés, ordre des arguments, exemples privilégiés, degré de prudence ou manière de réviser une conclusion.

Le résultat ne serait pas « la personne ». Ce serait une fonction construite pour certaines tâches à partir d'un corpus déterminé. Cette formulation

peut sembler froide. Elle protège pourtant contre une erreur majeure : attribuer au modèle tout ce que nous savons, aimons ou imaginons de l'individu.

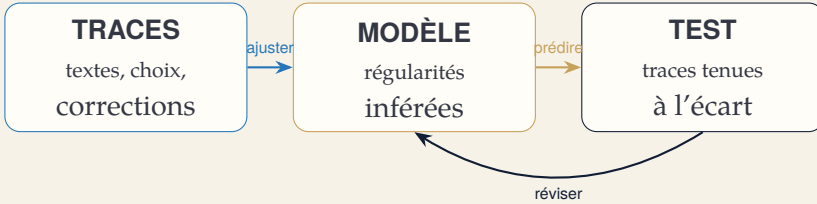


Figure. Un modèle de personne doit être évalué sur des productions non utilisées pour sa construction. La ressemblance n'est pas une preuve d'identité. Schéma conceptuel de l'auteur.

Trois situations documentaires

Comparer trois cas rend l'inférence mesurable.

Un document isolé. Une phrase apparaît dans une lettre. Plusieurs explications restent ouvertes. La même phrase, lue seule, peut être une formule de politesse, une position scientifique ou une concession tactique.

Plusieurs documents indépendants. Une lettre, une note de calcul et un texte public convergent. L'hypothèse gagne en solidité si les pièces ne répètent pas la même source sous des formes différentes. Cent copies d'un même récit ne valent pas cent preuves.

Documents contradictoires ou de périodes différentes. Deux formulations datées s'opposent. L'inférence doit alors nommer la période, exposer la tension ou s'abstenir. Une synthèse sans couture n'est pas une résolution : c'est une perte d'information.



Figure. Une inférence gagne en solidité lorsque les documents sont contextualisés, confrontés et testés sur de nouvelles pièces. Elle ne devient pas pour autant une certitude. Schéma de l'auteur.

Le genre du document compte autant que son contenu. Une phrase lue dans une lettre privée, un brouillon ou un texte public n'autorise pas les mêmes conclusions sur l'auteur. Attribuer une intention psychologique précise à Einstein à partir d'une seule notice de volume serait, ici, une faute de méthode.

Comment tester un modèle de personne

Le test le plus faible consiste à demander au modèle de reproduire ce qu'il a déjà vu. Un système pourrait restituer une formule ou imiter un passage sans avoir appris une régularité généralisable. Pour évaluer d'avantage, il faut réserver des productions qui n'ont pas servi à sa construction.

Supposons que nous disposions de lettres datées. Une partie forme le corpus de travail ; une autre reste cachée pendant le développement. Le modèle reçoit le contexte d'une lettre tenue à l'écart et doit prédire, non chaque mot exact, mais certains éléments définis à l'avance : position défendue, type d'argument, concept mobilisé ou degré de confiance.

Les critères doivent être fixés avant de regarder le résultat. Sinon, une réponse vague pourra toujours être déclarée « ressemblante » après coup. Il faut aussi comparer le système à des références : une règle simple, des lecteurs non spécialistes, des experts familiers du corpus. Un résultat n'est informatif que relativement à une base de comparaison.

Enfin, l'évaluation doit chercher les échecs. Que se passe-t-il lorsque le contexte change, lorsqu'une question est ambiguë ou lorsque deux périodes de la vie présentent des positions différentes ? Un modèle qui mélange ces périodes peut sembler cohérent tout en étant historiquement faux.

■ ÉTABLI

Évaluer sur des données tenues à l'écart réduit le risque de confondre restitution et généralisation. Ce principe ne suffit pas : les critères, les comparaisons et l'analyse des erreurs déterminent ce que le test permet réellement de conclure.

La personne change

Un modèle de personne contient souvent une hypothèse silencieuse : l'individu serait assez stable pour que le passé prévoie le futur. Cette hypo-

thèse est parfois raisonnable. Une préférence durable ou une compétence acquise crée des régularités. Elle devient trompeuse si elle efface l'apprentissage, les ruptures et les changements de contexte.

Les écrits d'un chercheur au début de sa carrière ne décrivent pas nécessairement ses positions ultérieures. Un désaccord peut modifier sa méthode. Une nouvelle preuve peut renverser une conclusion. Même le vocabulaire évolue. Un modèle entraîné sur toutes les périodes sans chronologie risque de produire une moyenne qui n'a jamais existé.

La temporalité doit donc faire partie des données. Il faut pouvoir demander : « selon les documents de quelle période ? » Une réponse qui reconnaît l'évolution est plus fidèle qu'une voix artificiellement uniforme.

Ce problème existe aussi pour les assistants personnalisés contemporains. Une préférence enregistrée hier peut devenir obsolète. Une information exacte peut cesser de l'être. La possibilité de consulter, corriger et supprimer une mémoire n'est pas seulement une fonction pratique ; elle répond à la nature changeante de la personne représentée.

Le miroir des évaluateurs

Pour juger si un modèle « ressemble » à quelqu'un, il faut des évaluateurs. Ceux-ci apportent leurs propres attentes. Un public peut reconnaître surtout les traits les plus célèbres d'une personnalité. Un spécialiste peut privilégier une période ou une interprétation historiographique. Un proche peut valoriser des aspects absents des archives publiques.

Le modèle risque alors de devenir un miroir des stéréotypes. S'il produit une phrase associée à l'image populaire d'Einstein, elle pourra sembler juste même si elle ne correspond à aucun document ou raisonnement attesté. Plus le personnage est célèbre, plus la reconnaissance intuitive doit être complétée par des critères documentaires.

Une évaluation robuste sépare donc plusieurs questions :

- le texte est-il compatible avec les sources de la période ?
- reproduit-il des régularités qui n'étaient pas explicitement fournies ?
- cite-t-il correctement la provenance lorsqu'il affirme un fait ?
- sait-il signaler qu'une réponse n'est pas déterminée par le corpus ?
- les experts s'accordent-ils, et sur quels critères ?

Le désaccord entre évaluateurs n'est pas seulement du bruit à éliminer. Il peut révéler que la notion de fidélité n'a pas été suffisamment définie.

Personnalité, identité, rôle

Un autre raccourci consiste à parler d'un « jumeau intellectuel ». Le terme évoque une copie globale. Dans la pratique, un système sera généralement meilleur pour certains rôles : retrouver des passages, simuler une forme de dialogue pédagogique, proposer une objection compatible avec un corpus ou imiter un style.

Ces rôles ne constituent pas une identité. L'identité d'une personne inclut une histoire vécue, des relations, un corps, des engagements et une capacité à répondre de ses actes. Un système qui génère des phrases ne devient pas le sujet historique auquel ces phrases ressemblent.

La distinction a des conséquences éthiques. Présenter une simulation comme une personne disparue peut créer une autorité trompeuse. L'utilisateur peut attribuer à l'individu des propos jamais tenus. Les désaccords historiques peuvent être lissés pour produire une voix plus agréable. Une interface doit donc annoncer la nature du système, ses sources, sa période de couverture et ses limites. Le vocabulaire choisi compte lui aussi : parler comme si le système était spontanément une personne peut faire oublier ce que ses mécanismes et ses évaluations établissent réellement [\[SRC-025\]](#).

■ ÉTABLI

Une similarité de style ou de réponse n'établit pas l'identité entre un modèle et une personne. Le statut de simulation, le corpus utilisé et la provenance des affirmations doivent rester visibles.

Ce qu'un modèle peut connaître

Nous pouvons maintenant donner une réponse conditionnelle à la question du chapitre. Oui, un modèle peut capturer des régularités réelles dans les productions d'une personne. Il peut parfois anticiper un choix, retrouver une structure récurrente ou proposer une réponse que des évaluateurs jugent compatible avec un corpus.

Non, cette réussite ne lui donne pas un accès total à la personne. Le modèle est limité par les traces disponibles, la manière dont elles ont été sélectionnées, les situations couvertes et l'objectif de l'entraînement. Il peut être précis et néanmoins muet sur ce qui n'a jamais été exprimé. Il peut être convaincant et néanmoins faux hors de sa distribution de test.

La bonne question n'est donc pas : « le système connaît-il vraiment cette personne ? » posée sans autre précision. Il faut demander :

Quelles caractéristiques des productions peut-il estimer, avec quelle fiabilité, dans quelles conditions, et quelles observations pourraient montrer qu'il se trompe ?

Cette formulation transforme une impression en programme de recherche.

Le seuil de l'expérience de pensée

Revenons à Einstein. Nous disposons d'un corpus historique riche mais incomplet, édité par des institutions et traversé par plusieurs langues, périodes et genres [SRC-001] [SRC-032] [SRC-036]. Nous pourrions imaginer un système qui explore ce corpus et génère des réponses. Pour savoir ce qu'il représente, il faudrait séparer les documents retrouvés, les régularités apprises et les attentes des évaluateurs.

Nous devrions surtout préserver la frontière entre deux Einstein. L'Einstein historique apparaît dans des documents attribués et contextualisés. L'Einstein imaginé sert à poser des questions contemporaines. Une réponse générée n'appartient ni à l'un ni à l'autre : elle est la sortie d'un système construit par nous.

Cette frontière ouvre la partie suivante. Nous pouvons maintenant examiner l'expérience de pensée en détail : quel assistant construirions-nous, quelles traces lui donnerions-nous et comment éviterions-nous de prendre sa voix pour une résurrection ?

À RETENIR

- 01 Après la transformation des traces vient l'inférence sur l'auteur.
- 02 Un document isolé, des pièces indépendantes et des contradictions n'ont pas le même pouvoir.
- 03 Multiplier les copies d'une même source n'équivaut pas à multiplier les preuves.
- 04 Un modèle de personne doit annoncer tâche, période et corpus.
- 05 La similarité d'une sortie n'établit ni identité ni autorité historique.

Sources mobilisées

- **SRC-001** - **Einstein Papers Project**. Sert à caractériser le corpus historique sans attribuer de réaction fictive à Einstein.
- **SRC-012** - **Park et al., Generative Agents**. Métadonnées, résumé et ex-

traits de méthode examinés ; exemple d'architecture d'agents, non modèle fidèle d'une personne réelle.

- **SRC-017 - Conway et Pleydell-Pearce.** Mobilisé pour la relation entre mémoire autobiographique, buts présents et représentation du soi.
- **SRC-022 - Frith et Frith, mentalisation.** Métadonnées, résumé et extraits examinés ; cadre pour l'inférence sur les états mentaux d'autrui.
- **SRC-023 - Rabinowitz et al., Machine Theory of Mind.** Notice PMLR, résumé et extraits d'introduction examinés ; exemple limité à des agents artificiels en mondes-grilles.
- **SRC-024 - Dennett, stratégie intentionnelle.** Notice et extraits examinés ; position philosophique, non résultat expérimental.
- **SRC-025 - Shanahan, langage employé à propos des LLM.** Métadonnées, résumé et extraits de conclusion examinés ; utilisé pour prévenir l'anthropomorphisme non justifié.
- **SRC-032 - Einstein Papers Project, histoire du projet.** Page institutionnelle intégralement consultée ; elle documente l'histoire éditoriale, non chaque document.
- **SRC-036 - Collected Papers, volume 5.** Notice éditoriale officielle consultée ; elle précise la portée générale d'un corpus de correspondance.
- **SRC-028 - Youyou, Kosinski et Stillwell, jugements de personnalité.** Notice DOI et résumé PubMed examinés ; résultats limités à un échantillon, une plateforme et des cibles psychométriques.
- **SRC-029 - Kosinski, Stillwell et Graepel, traces numériques.** Notice DOI et résumé PubMed examinés ; prédictions probabilistes, non certitudes individuelles.
- **SRC-030 - Argyle et al., simulation d'échantillons humains.** Notice et résumé Cambridge examinés ; distributions collectives, non fidélité personnelle.
- **SRC-031 - Aher, Arriaga et Kalai, réplication d'études humaines.** Notice PMLR et résumé examinés ; le PDF n'est pas revendiqué comme intégralement lu.

4 · Einstein en 2026

Le scénario peut éclairer une méthode, à condition de ne jamais faire passer l'interface imaginée pour une archive.

■ EXPÉRIENCE DE PENSÉE

Einstein en 2026 est une fiction méthodologique. Aucun geste, aucune question et aucune réponse décrits dans cette partie ne sont des faits historiques.

Deux présences dans la même pièce

Imaginons de nouveau la pièce de travail. L'écran est allumé. Le carnet est ouvert. Cette fois, la question n'est plus de savoir ce qu'est une pensée en général. Elle est de savoir quel Einstein nous aurions réellement placé devant cet écran.

Il y en a au moins deux, et ils ne doivent pas se recouvrir. Le premier est historique. Il est mort en 1955. Nous le connaissons par des documents édités, datés et commentés : lettres, articles, notes et calculs [\[SRC-001\]](#) [\[SRC-032\]](#). Le second est un personnage de méthode. Il sert à demander ce qu'un assistant contemporain pourrait faire avec ces documents. Il ne peut pas témoigner à la place du premier.

Cette séparation paraît simple. Elle est pourtant la condition de toute la partie. Si nous la relâchons, une réponse fluide pourra être lue comme une citation, et une citation réelle pourra être noyée dans une scène inventée.

Ce que les archives permettent déjà de voir

Les notices des *Collected Papers* donnent une idée de la diversité du matériau, sans remplacer les pièces elles-mêmes. Le volume des années de jeunesse signale une correspondance précoce et un intérêt pour l'électrodynamique des corps en mouvement [\[SRC-033\]](#). Le volume des écrits suisses rassemble des travaux scientifiques de 1900 à 1909, dont les articles de 1905 [\[SRC-034\]](#). Le volume 4 indique des notes et des calculs liés à la progression

vers une théorie relativiste de la gravitation [SRC-035]. Le volume 5 décrit plus de cinq cents lettres pour 1902-1914 [SRC-036].

Ces indications autorisent une conclusion limitée : le corpus n'est pas une seule voix publique. Il comprend des genres, des périodes et des destinataires différents. Elles n'autorisent pas à raconter le contenu d'une lettre précise ou le chemin exact d'un calcul tant que la pièce n'a pas été consultée.

■ ÉTABLI

Les notices officielles décrivent la portée de volumes et de catégories documentaires. Elles ne remplacent pas la lecture des documents numérotés.

Encadré documentaire - le corpus hypothétique

Pour l'expérience de pensée, l'assistant n'aurait pas « Einstein ». Il aurait un corpus édité, avec des catégories distinctes :

- **Textes scientifiques** publiés ou préparés dans une période nommée.
- **Correspondance** datée, avec destinataires et langues variables.
- **Notes et calculs** signalés dans certaines notices de volumes.
- **Commentaires éditoriaux** qui ne sont pas la voix d'Einstein.
- **Traductions** et présentations ultérieures, à étiqueter comme telles.

L'Einstein Papers Project fournit une source institutionnelle pour vérifier la portée des archives et les caractéristiques des volumes [SRC-001] [SRC-032]. Un accès à une collection éditée n'équivaut pas à la possession de toutes les traces de la vie intellectuelle d'Einstein. Des absences, des pièces non numérisées et des choix éditoriaux demeurent.

La question concrète de l'enquête

Avant de parler d'un modèle de pensée, l'assistant reçoit une tâche étroite :

À partir d'un ensemble de documents datés d'une période définie, peut-on

décrire comment une formulation scientifique évolue entre deux versions

sans inventer la cause du changement ?

Cette question prolonge l'exemple du chapitre 1. Elle n'exige pas de reconstituer une conscience. Elle exige de distinguer observation documentaire, inférence et inconnu - puis de citer les pièces.

Le mythe est lui aussi une trace

Un assistant nourri seulement d'anecdotes célèbres n'apprendrait pas nécessairement la méthode documentée. John D. Norton soutient, dans l'introduction d'un article de 2016, que les découvertes de 1905 ont été précédées d'enquêtes longues, souvent ordinaires dans leur labeur, et que plusieurs récits populaires les réduisent à des raccourcis [SRC-040]. Il annonce quatre mythes : la pensée enfantine, la pensée purement opérationnelle, le simple plaisir de briser les règles et le pouvoir d'une expérience unique.

Cette thèse est celle de l'article, établie à partir de son résumé et de son introduction. Les études de cas qui suivent n'ont pas été examinées intégralement ici. Elle suffit néanmoins à poser une exigence : un modèle d'Einstein devrait pouvoir être confronté à des documents de travail, pas seulement à l'image du génie instantané.

Le risque est concret. Si les données les plus disponibles sont des biographies simplifiées, des posters ou des phrases détachées, le système reproduira d'abord cette circulation culturelle. Il pourra sembler convaincant précisément parce qu'il confirme une attente déjà répandue.

Ce que la scène de 2026 ajoute

Dans la fiction, l'Einstein imaginé pourrait interroger l'assistant, corriger une réponse ou laisser une question sans suite. Ces gestes créeraient de nouvelles traces. Elles n'appartiendraient pas aux archives historiques. Elles appartiendraient au scénario et au système qui y répond.

Il faudrait donc tenir deux journaux. L'un recense les documents historiques et leur provenance. L'autre recense les interactions imaginées ou simulées, avec la date de la simulation, le modèle utilisé et les textes fournis dans le contexte. Mélanger ces journaux produirait une archive fictive.

Une expérience de pensée reste utile parce qu'elle organise un scénario pour examiner des conséquences [SRC-038]. Ici, la conséquence à examiner est pratique : quelles catégories de documents, quelles métadonnées et quels tests seraient nécessaires avant de parler d'un assistant intellectuel ?

Ce qui resterait inconnu

Même un corpus très riche laisserait des zones fermées. Nous ne saurions pas quelles hésitations n'ont jamais été écrites. Nous ne saurions pas quelle formulation a été choisie pour un destinataire plutôt que parce qu'elle semblait la plus exacte. Nous ne saurions pas comment Einstein aurait réagi à une interface qui n'existait pas.

Ces inconnues ne rendent pas le projet vain. Elles en fixent le bord. Le chapitre suivant peut donc quitter le personnage et décrire l'outil : non pas ce qu'Einstein aurait dit, mais quels mécanismes distincts un assistant devrait exposer pour que ses réponses restent interprétables.

À RETENIR

- 01 Einstein en 2026 est un dispositif fictif, distinct de l'Einstein documenté.
 - 02 Le corpus hypothétique doit séparer genres, périodes et commentaires éditoriaux.
 - 03 La question concrète porte sur l'évolution documentaire d'une formulation, pas sur une résurrection.
 - 04 Les mythes de découverte peuvent devenir des données plus visibles que les dossiers.
 - 05 Le chapitre suivant décrit l'architecture de l'outil.
-

Sources mobilisées

- **SRC-001 - Einstein Papers Project.** Portée institutionnelle consultée ; elle ne vaut pas inventaire de chaque document.
- **SRC-032 - Histoire du projet.** Page institutionnelle intégralement consultée.
- **SRC-033 - Volume 1.** Notice éditoriale officielle ; documents individuels non consultés.
- **SRC-034 - Volume 2.** Notice éditoriale officielle ; la date commerciale peut différer de la date d'édition et doit rester signalée.
- **SRC-035 - Volume 4.** Notice éditoriale officielle ; manuscrits non parcourus page à page.
- **SRC-036 - Volume 5.** Notice éditoriale officielle ; aucune lettre individuelle n'est citée à partir de cette seule notice.

- **SRC-038 - Expériences de pensée.** Entrée académique consultée intégralement pour encadrer le statut du scénario.
- **SRC-040 - Norton, mythes de découverte.** Résumé et introduction examinés; les études de cas ultérieures ne sont pas revendiquées comme lues intégralement.

5 · Le chatbot personnel d'Einstein

Un assistant intellectuel n'est pas une personne réduite à une fenêtre : c'est un assemblage de corpus, de contexte, de mémoire et de recherche.

■ EXPÉRIENCE DE PENSÉE

Le chatbot décrit ici n'a pas été construit. Le chapitre spécifie ce qu'il faudrait séparer pour qu'une telle construction reste intelligible.

Une fenêtre, quatre opérations

Le chapitre précédent a fixé le personnage. Celui-ci fixe l'outil. Dans la scène fictive, une seule fenêtre recevrait la question. Cette simplicité d'interface ne doit pas décider de l'architecture.

Quatre opérations restent distinctes. Les **paramètres** viennent d'un entraînement antérieur. Le **contexte** rassemble ce qui est fourni pour la réponse en cours. Une **mémoire de produit** peut conserver certaines informations et les réinjecter plus tard, selon des réglages. Une **recherche externe** peut retrouver des documents et les placer dans ce contexte.

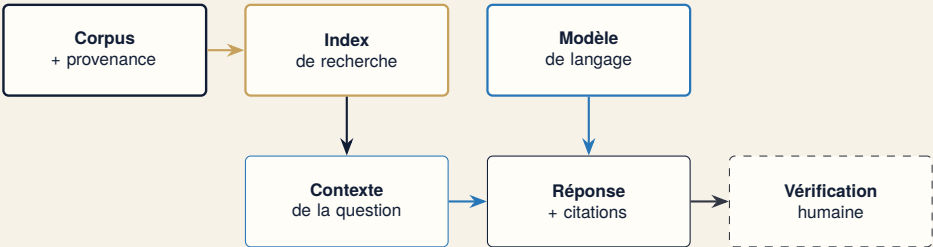
La documentation de ChatGPT, consultée le 9 octobre 2026, sépare les contrôles de mémoire des contrôles concernant l'usage de conversations pour améliorer les modèles [SRC-006] [SRC-007]. Cette séparation est une politique de produit, datée et modifiable. Elle n'est pas une loi de tous les systèmes.

■ ÉTABLI

Dans les documents consultés, mémoriser une préférence, retrouver un document et ajuster des poids ne sont pas la même opération.

Architecture fonctionnelle hypothétique

Pour rendre l'enquête visible, l'assistant imaginé peut être décrit comme une chaîne de fonctions, non comme une personne dans une fenêtre.



Architecture proposée pour l'expérience de pensée. Ce n'est ni ChatGPT ni un système historique réel.

Figure. Architecture fonctionnelle hypothétique d'un assistant historique. Proposition pour l'expérience de pensée, non description d'un produit existant. Schéma de l'auteur.

Fonction	Question	Limite
Recherche	Où est le document pertinent ?	Le passage peut être inadéquat
Mise en contexte	Quelle période ou situation ?	Le contexte peut être incomplet
Comparaison	Quelles différences entre pièces ?	La cause n'est pas automatique
Génération	Comment formuler une réponse ?	Une phrase peut dépasser les preuves
Vérification	Les sources étayent-elles l'affirmation ?	La vérification doit elle-même être contrôlée

Tableau. Fonctions d'un assistant historique hypothétique. Architecture proposée, non un système réellement construit pour Einstein.

Cette architecture est une proposition pour l'expérience de pensée. Elle n'identifie pas ChatGPT. Elle ne décrit pas un système historique réellement construit pour Einstein. Elle sert à montrer ce qu'il faudrait exposer pour qu'une réponse reste interprétable : corpus, recherche, contexte, génération, citations, vérification.

Ajuster les poids n'est pas discuter

La différence devient plus nette si l'on regarde un travail d'entraînement. La documentation OpenAI de l'ajustement supervisé décrit une suite d'étapes : constituer des exemples, les téléverser, créer un travail d'entraînement, puis évaluer le modèle obtenu. Elle présente ce travail comme ce qui met à jour les poids à partir des exemples [SRC-039]. La même page indique que la disponibilité de cette plateforme se réduit. Il faut donc dater l'information et ne pas la transformer en description permanente de ChatGPT.

Le résultat de Brown et ses collègues va dans le même sens pour une autre situation : les tâches few-shot évaluées dans leur article sont réalisées

sans mise à jour de gradient pendant ces tâches [SRC-009]. Un exemple placé dans une consigne peut changer la réponse présente. Il ne constitue pas, par cela seul, un nouvel entraînement du modèle général.

Un assistant personnel pourrait utiliser les deux régimes, mais il devrait le dire. « Cette réponse s’appuie sur des passages retrouvés » n’a pas le même sens que « les poids ont été ajustés sur un corpus historique ».

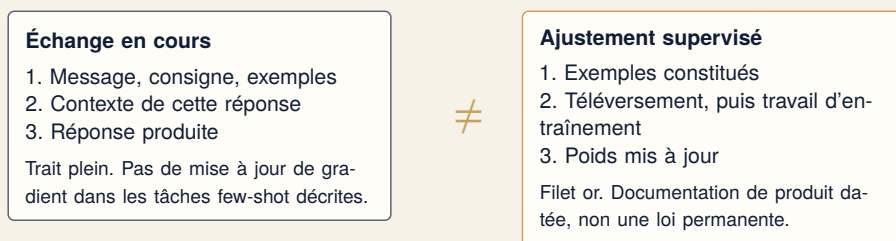


Figure. À gauche, le message change le contexte de la réponse. À droite, un travail d’entraînement met à jour des poids. Schéma explicatif. [SRC-009], notice et résumé. [SRC-039], page consultée le 9 octobre 2026.

Le corpus devrait être décrit avant d’être interrogé

Avant toute question, il faudrait une fiche de corpus. Elle indiquerait la période, les langues, les genres et le statut de chaque pièce : article publié, lettre, note de calcul, traduction ou commentaire éditorial. Elle distinguerait aussi le texte d’Einstein des introductions écrites par les éditeurs.

Les notices déjà consultées montrent que ces genres existent et qu’ils ne sont pas interchangeables [SRC-033] [SRC-034] [SRC-035] [SRC-036]. Une lettre à un proche et un mémoire scientifique ne fournissent pas le même type d’indice. Une note de calcul peut documenter une tentative. Un texte public peut simplifier, pour un lecteur, un résultat longuement préparé [SRC-040].

Les métadonnées minimales seraient donc la date, l’auteur attribué, le destinataire lorsqu’il existe, la langue, le genre, l’édition utilisée et l’état de vérification. Sans elles, une recherche pourrait rapprocher des phrases que rien, historiquement, ne rapproche.

Personnaliser n’est pas ressusciter

Une mémoire de produit peut retenir qu’un utilisateur préfère des réponses courtes ou souhaite toujours les références. Cette fonction est utile. Elle produit un profil d’usage. Elle ne produit pas la mémoire autobiographique de la personne dont on étudie les archives.

Les agents génératifs de Park et ses collègues combinent un historique en langage naturel, une récupération, une réflexion et une planification. Leur démonstration concerne vingt-cinq agents dans un environnement contrôlé [SRC-012]. Elle montre qu'une architecture peut rendre un comportement plus suivi. Elle ne montre pas qu'un agent devient la personne dont il imite certains rôles.

Pour Einstein, la règle devrait être visible dans l'interface elle-même : corpus couvert, période couverte, documents retrouvés, parts générées et niveau d'incertitude. Une voix continue peut masquer ces couches. Le chapitre suivant demande comment mesurer si, malgré cette transparence, le système apprend autre chose qu'une imitation de surface.

À RETENIR

- 01 L'architecture hypothétique sépare corpus, recherche, modèle, contexte et vérification.
- 02 Chaque fonction a une limite : retrouver n'est pas prouver.
- 03 La mise à jour des poids reste distincte d'un échange en cours.
- 04 La personnalisation d'un produit n'est pas la mémoire vécue d'une personne.
- 05 Le chapitre suivant transforme cette architecture en protocole d'évaluation.

Sources mobilisées

- **SRC-006 - Contrôles de données OpenAI.** Page officielle consultée le 9 octobre 2026.
- **SRC-007 - Mémoire OpenAI.** Page officielle consultée le 9 octobre 2026 ; fonctions variables selon l'offre.
- **SRC-009 - Brown et al.** Notice et résumé ; pas de mise à jour de gradient pendant les tâches few-shot décrites.
- **SRC-012 - Park et al.** Résumé et extraits de méthode ; vingt-cinq agents fictifs.
- **SRC-033, SRC-034, SRC-035 et SRC-036 - Notices des volumes.** Catégories documentaires, pas contenu de chaque pièce.
- **SRC-039 - Ajustement supervisé OpenAI.** Page consultée le 9 octobre 2026 ; disponibilité de la plateforme en évolution.

- **SRC-040 - Norton.** Introduction utilisée seulement pour rappeler que les textes publics peuvent simplifier un long travail.

6 · Le modèle d'Einstein

Un modèle n'a de portée que si l'on peut dire quelle tâche il réalise, sur quelle période et contre quelles pièces tenues à l'écart.

■ EXPÉRIENCE DE PENSÉE

Aucun modèle d'Einstein n'est évalué dans ce chapitre. Le protocole décrit est une proposition de recherche à construire, non un résultat obtenu. Aucun score n'est inventé.

Nommer la tâche avant le système

« Modéliser Einstein » est une formule trop large pour être testée. Elle mélange le style, les concepts, les changements d'avis, les relations et l'expérience vécue. Un protocole doit choisir une tâche étroite.

Trois tâches candidates, exclusives l'une de l'autre pour un premier essai :

1. **Genre documentaire.** À partir d'un extrait daté, estimer s'il s'agit plutôt d'une note de calcul, d'un texte public ou d'une lettre, selon les catégories déjà visibles dans les notices des volumes [\[SRC-033\]](#) [\[SRC-035\]](#) [\[SRC-036\]](#). 2. **Cohérence de période.** Dire si une formulation proposée est compatible avec le dossier d'une période nommée, sans la présenter comme une phrase d'Einstein. 3. **Provenance.** Distinguer une affirmation documentée dans le corpus fourni d'une continuation inventée.

La tâche doit annoncer sa période. Les notices montrent déjà que 1900-1909 et 1912-1914 ne présentent pas les mêmes dossiers [\[SRC-034\]](#) [\[SRC-035\]](#). Un système qui répond avec la voix d'une période à une question posée dans une autre peut paraître cohérent tout en étant historiquement faux.

■ HYPOTHÈSE

Un système construit sur des traces sélectionnées pourrait reproduire certaines régularités de ces traces sans reproduire l'expérience subjective d'Einstein. Cette possibilité reste à tester.

Protocole proposé - huit cases

Voici la grille minimale. Elle n’a pas été exécutée dans ce livre.

1. **Question.** Peut-on prédire une propriété documentaire définie sans reconstruire une personne ? 2. **Corpus.** Édition savante identifiable ; genres séparés ; commentaires d’éditeurs étiquetés comme tels [SRC-001]. 3. **Période.** Une seule fenêtre chronologique par expérience. 4. **Séparation.** Pièces de préparation, pièces de développement, pièces d’évaluation tenues à l’écart. 5. **Configurations.** (a) recherche documentaire seule ; (b) contexte few-shot sans mise à jour de gradient pendant la tâche, régime décrit pour GPT-3 [SRC-009] ; (c) éventuel ajustement de poids sur un jeu d’exemples, régime distinct [SRC-039]. 6. **Base de comparaison.** Lecteurs humains disposant du même dossier, et une règle simple de fréquence de termes. 7. **Critères fixés avant lecture.** Exactitude sur la tâche, citation de provenance, stabilité sous reformulation, taux d’invention de pièces absentes. 8. **Publication des échecs.** Changement de langue, de destinataire, de période, confusion éditeur / auteur.

Cette séparation reprend la distinction entre retrouver un document et ajuster des poids [SRC-004] [SRC-039]. Un système de recherche peut citer la pièce tenue à l’écart s’il y a accès au moment du test. Ce n’est pas une généralisation. La généralisation commence lorsque la pièce n’a servi ni à l’index de réponse, ni à l’ajustement.

Étape	Exigence	Échec typique
Tâche	Une propriété documentaire précise	« Modéliser Einstein » sans critère
Période	Une fenêtre chronologique unique	Mélange 1905 / 1913
Hold-out	Pièces tenues hors construction	Fuite via recherche ou contexte
Critères	Exactitude, provenance, stabilité	Juger une « voix juste »

Tableau. Grille minimale du protocole proposé au chapitre 6. Proposition de recherche, non un résultat obtenu.

Une seule tâche principale pour le premier essai

Pour éviter de diluer le protocole, la première expérience retient la **fidélité documentaire** : le système sait-il répondre à une question sur une période donnée en citant les documents qui étayaient ses affirmations ?

Cette cible est plus contrôlable que « reproduire la pensée d’Einstein ». Elle reprend la question du chapitre 4 sur l’évolution d’une formulation. Elle exige des citations vérifiables, une période nommée et une capacité d’abstention lorsque le corpus ne suffit pas.

Séparation des données

Trois ensembles doivent rester distincts avant toute lecture des résultats.

- 1. **Construction.** Pièces utilisées pour constituer l’index documentaire ou, le cas échéant, pour un ajustement de poids.
- 2. **Développement.** Pièces servant à régler des consignes ou des seuils, sans entrer dans le score final.
- 3. **Évaluation.** Pièces tenues à l’écart. Une pièce accessible via la recherche au moment du test ne peut pas être présentée comme preuve de généralisation indépendante à cette pièce.

Critère	Question	Échec typique
Exactitude	Correct selon les pièces ?	Invention plausible
Provenance	Les citations existent-elles ?	Fausse référence
Temporalité	La période est-elle respectée ?	Mélange de dates
Robustesse	Résiste à une reformulation ?	Succès de wording
Abstention	Signale-t-il l'absence de preuves ?	Réponse forcée
Contamination	Fuite des pièces de test ?	Fausse généralisation

Tableau. Grille d’évaluation proposée pour le protocole du chapitre 6. Critères à fixer avant l’examen des résultats. Aucun score n’a été obtenu ici.

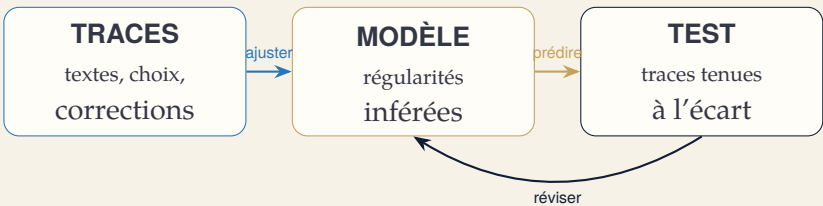


Figure. Un modèle de personne doit être évalué sur des productions non utilisées pour sa construction. La ressemblance n’est pas une preuve d’identité. Schéma conceptuel de l’auteur.

Catégories d’erreurs à prévoir avant le test

Une réponse peut échouer pour des raisons différentes. Les nommer à l’avance empêche de traiter tous les échecs comme un seul verdict.

- **Mauvais document** : citation d’une pièce hors sujet ou hors période.
- **Confusion temporelle** : mélange de deux moments du corpus.
- **Interprétation présentée comme fait** : une inférence n’est plus étiquetée.
- **Synthèse lissante** : une contradiction documentée disparaît.

- **Invention de provenance** : référence élégante absente des archives.
- **Contamination** : la réponse s'appuie sur une pièce réservée au test.

Ce qui compterait comme réussite, et comme échec

Réussite limitée. Sur la tâche de fidélité documentaire, le système cite correctement, respecte la période, s'abstient lorsque les pièces manquent, et dépasse une règle simple de fréquence sans inventer de lettres.

Échec informatif. Le système invente une provenance, mélange 1905 et 1913, attribue à Einstein un commentaire d'éditeur, ou réussit seulement lorsqu'on lui montre la réponse dans le contexte.

Ce qui ne compterait jamais comme preuve. Une voix « qui sonne juste ». Une reconnaissance imaginaire. Une impression de compréhension. Le chapitre suivant explique pourquoi ces jugements, même sincères, ne remplacent pas la grille ci-dessus.

L'échelle donnée par les simulations existantes

Quelques résultats donnent une échelle, pas un modèle d'Einstein. Rabinowitz et ses collègues ont prédit des agents artificiels dans des mondes-grilles [\[SRC-023\]](#). Park et ses collègues ont obtenu des comportements jugés suivis chez des agents fictifs [\[SRC-012\]](#). Argyle et ses collègues ont montré que GPT-3, conditionné par des profils, pouvait reproduire certaines distributions de réponses de groupes américains [\[SRC-030\]](#). Aher, Arriaga et Kalai ont reproduit certains résultats d'études et observé aussi une distorsion de sur-précision [\[SRC-031\]](#).

Aucun de ces travaux n'étudie Einstein. Leur leçon commune est plus étroite : une simulation peut être utile à une tâche et rester infidèle à une personne. Elle peut même être trop régulière, là où les réponses humaines se dispersent.

Les prédictions de traits à partir de traces numériques appellent la même réserve. Dans des données Facebook précises, certains attributs ou jugements de personnalité ont été prédits statistiquement [\[SRC-028\]](#) [\[SRC-029\]](#). La performance dépend de la population, de la plateforme et de la cible. Elle ne reconstitue pas une biographie intellectuelle.

Ce que ce chapitre ajoute

Les chapitres précédents ont fixé le personnage et l'outil. Celui-ci fixe le banc d'essai. Sans tâche, période, hold-out et critères d'échec, « le modèle

d'Einstein » reste un slogan. Avec eux, il devient une hypothèse que l'on peut, un jour, mettre en défaut.

À RETENIR

- 01 La tâche, la période et le corpus doivent être nommés avant l'évaluation.
 - 02 Les pièces de test doivent rester hors de la construction et de la recherche utilisée pour répondre.
 - 03 Une voix convaincante n'est pas un critère.
 - 04 Les simulations existantes portent sur d'autres agents ou groupes.
 - 05 Le protocole de ce chapitre n'a pas été exécuté ici.
-

Sources mobilisées

- **SRC-001** - **Einstein Papers Project**. Portée d'une édition, non lecture pièce à pièce.
- **SRC-004** - **Lewis et al.** Résumé ; mémoire paramétrique et récupération.
- **SRC-009** - **Brown et al.** Notice et résumé ; few-shot sans mise à jour de gradient pendant les tâches décrites.
- **SRC-012**, **SRC-023**, **SRC-030**, **SRC-031**. Simulations limitées ; aucune sur un corpus einsteinien.
- **SRC-028** et **SRC-029**. Prédictions statistiques dans des données numériques spécifiques.
- **SRC-033** à **SRC-036**. Notices de volumes pour genres et périodes.
- **SRC-039** - **Ajustement supervisé**. Page consultée le 9 octobre 2026.

7 · Einstein aurait-il reconnu son propre modèle ?

La reconnaissance imaginaire peut révéler nos attentes. Elle ne peut pas servir de preuve historique.

■ EXPÉRIENCE DE PENSÉE

Aucune reconnaissance par Einstein n'est envisageable comme fait. La question sert à examiner la valeur et la faiblesse de ce critère.

Le juge que nous aimerions avoir

Après un test réussi, une question revient presque naturellement : Einstein aurait-il reconnu cette réponse comme sienne ? Elle est puissante parce qu'elle promet un témoin ultime. Elle est aussi inutilisable comme preuve. Le témoin historique ne peut pas comparer la sortie du système à son intention. Le faire répondre dans la fiction ne créerait qu'un second texte généré.

La reconnaissance, même lorsqu'elle concerne une personne vivante, n'est pas un instrument simple. Une personne peut se reconnaître dans un portrait flatteur et rejeter un portrait exact. Elle peut oublier une position ancienne ou préférer la formulation qu'elle donnerait aujourd'hui. Pour un corpus posthume, cette difficulté devient une impossibilité pratique.

■ INCONNU

Aucun protocole ne permet de faire reconnaître par Einstein historique un modèle construit après 1955. La question doit rester un outil critique, non un test.

Cinq sens de la fidélité

Le mot « reconnaître » masque plusieurs jugements. Ils ne sont pas équivalents.

Fidélité	Ce qui serait évalué	Ce que cela ne prouve pas
Stylistique	Vocabulaire et forme	Fidélité historique profonde
Documentaire	Exactitude des citations	Compréhension intégrale
Historique	Cohérence période et corpus	Expérience vécue
Scientifique	Validité d'une démonstration	Démarche historique d'Einstein
Comportementale	Régularités sur plusieurs tâches	Identité personnelle

Tableau. Cinq sens de la fidélité. Les confondre transforme une reconnaissance intuitive en preuve.

Une réponse peut ressembler au style attendu sans être cohérente avec les documents. Elle peut citer correctement sans restituer une démarche historique. Elle peut résoudre un problème scientifique selon des critères définis sans reproduire le chemin qu'Einstein aurait effectivement suivi. Le protocole du chapitre 6 devient nécessaire précisément parce que ces jugements se confondent trop facilement.

Ressemblance et stéréotype

Une sortie peut être reconnue par des lecteurs pour de mauvaises raisons. Elle confirme l'image qu'ils attendent : une illumination soudaine, une règle brisée avec légèreté, une expérience unique qui tiendrait lieu d'enquête. Norton présente précisément ces récits comme des simplifications capables d'éclipser des années de travail documenté [SRC-040]. Un modèle entraîné sur les versions les plus répétées de ces récits aurait de bonnes chances de les restituer. Son succès mesurerait alors la force du stéréotype, pas la fidélité aux dossiers.

Il faut donc demander qui reconnaît, à partir de quelles sources, et en concurrence avec quelle autre réponse. Des historiens disposant des pièces tenues à l'écart n'ont pas le même jugement que des lecteurs nourris de l'image publique. Leur désaccord n'est pas un bruit à effacer. Il indique que le mot « ressemblance » recouvre plusieurs cibles.

Les travaux sur la simulation de groupes rappellent qu'une distribution plausible peut encore être trop lisse ou trop précise [SRC-030] [SRC-031]. Appliquée à une personne, cette leçon interdit de prendre l'aisance pour un aveu d'identité.

Ce qu'un désaccord apprendrait

Dans le scénario fictif, l'Einstein imaginé pourrait rejeter une réponse que les tests jugent fidèle à ses textes anciens. Ce rejet ne prouverait pas que les textes ont été mal lus. Il pourrait montrer qu'une personne change, qu'elle s'adresse à un public, ou que le système a confondu une période

avec une autre. Il pourrait aussi ne rien montrer du tout, puisque le rejet lui-même serait inventé.

La seule version exploitable du désaccord est documentaire. Deux pièces datées peuvent soutenir des formulations incompatibles. Le modèle doit alors les présenter ensemble, ou signaler qu'il choisit une période. S'il fabrique une synthèse sans couture, il ne résout pas la tension. Il la dissimule.

C'est aussi ici que le vocabulaire devient un risque. Décrire le système comme s'il reconnaissait, regrettait ou pensait peut faire oublier les mécanismes évalués [SRC-025]. Le livre peut employer ces verbes dans une expérience de pensée clairement annoncée. Il ne peut pas les laisser glisser vers un compte rendu historique.

L'acquis de la partie

Au terme de ces quatre chapitres, une réponse conditionnelle est possible. Un assistant pourrait organiser un corpus, retrouver des passages, distinguer des régimes techniques et soumettre certaines régularités à des tests. Il pourrait échouer de manière instructive. Il ne pourrait ni faire parler l'Einstein historique, ni transformer une reconnaissance imaginaire en certificat.

Si l'on ne peut pas demander à Einstein historique de reconnaître son modèle, **quels critères indépendants permettent d'en mesurer la fidélité ?** Le chapitre 6 a préparé les outils : exactitude, provenance, temporalité, robustesse, abstention, contamination. Le présent chapitre explique pourquoi ils sont indispensables. Sans eux, le titre du livre se refermerait sur une impression.

Cette limite ouvre la partie suivante. Pour comprendre ce que cet assistant ferait réellement, il ne suffit plus de décrire un corpus particulier. Il faut expliquer comment un modèle apprend à partir de données, ce que ses paramètres retiennent et pourquoi une réponse n'est pas une fiche extraite d'une base.

À RETENIR

- 01 La reconnaissance par Einstein historique est impossible à recueillir.
- 02 Stylistique, documentaire, historique, scientifique et comportementale ne se confondent pas.
- 03 Une ressemblance peut venir d'un stéréotype plus que d'un dossier.

04 Les critères du chapitre 6 remplacent le juge impossible.

05 La suite explique les mécanismes techniques qui rendent ces critères nécessaires.

Sources mobilisées

- **SRC-025 - Shanahan.** Résumé et extraits de conclusion ; prudence sur le vocabulaire mental appliqué aux modèles.
- **SRC-030 et SRC-031.** Résultats de simulation collective et distorsions ; analogie limitée, non transfert direct à Einstein.
- **SRC-040 - Norton.** Introduction consultée ; les quatre mythes nommés servent à identifier des stéréotypes de découverte, non à résumer toutes les études de cas de l'article.

8 · Des données aux représentations

Un dossier d'exemples n'est pas le modèle. L'entraînement transforme des exemples en poids, selon un objectif.

Le dossier n'est pas la compétence

On peut poser mille pages sur une table. Elles restent mille pages. Quelqu'un qui les a lues peut, ensuite, répondre à une question sans rouvrir chaque feuille. Le dossier et la compétence acquise ne sont pas le même objet.

Un système d'apprentissage automatique sépare, lui aussi, les exemples et ce qui en est retenu. Les exemples servent pendant une phase de préparation. Le résultat utilisable est un ensemble de valeurs numériques, les paramètres, ou poids. Une fois cette phase terminée, répondre à une question nouvelle n'est pas rouvrir le dossier page à page.



Schéma explicatif : le dossier d'exemples n'est pas le modèle obtenu.

Figure. Des exemples sélectionnés et un objectif d'apprentissage alimentent un travail d'entraînement. Le résultat utilisable est un ensemble de poids, non le dossier lui-même. Schéma explicatif de l'auteur. [SRC-010] [SRC-039], niveaux d'examen du registre.

Cette image a une limite. Une personne peut justifier une réponse en retrouvant le passage. Un modèle peut produire une suite de mots sans pouvoir montrer la page dont elle viendrait. La ressemblance pédagogique s'arrête là.

Ce que le schéma permet de vérifier

Trois questions suivent immédiatement. Quels exemples ont été retenus ? Quel objectif a guidé l'ajustement ? Les poids obtenus ont-ils été évalués sur des pièces tenues à l'écart ? Sans ces trois réponses, « le modèle a appris » reste une formule vague.

Ajuster, ce n'est pas converser

La documentation d'ajustement supervisé consultée le 9 octobre 2026 décrit une opération précise. On constitue des exemples. On les téléverse. Un travail d'entraînement met ensuite à jour les poids [SRC-039]. La même page indique que la disponibilité de cette plateforme se réduit. La description doit donc rester datée.

Le résumé de l'article d'Ouyang et de ses collègues présente une autre suite, pour les modèles InstructGPT : un ajustement sur des démonstrations, puis un apprentissage à partir de préférences humaines [SRC-010]. Ce résumé porte sur une famille de modèles et une distribution de consignes. Il signale aussi que des erreurs simples peuvent persister. Il ne dit pas qu'une conversation ordinaire, plus tard, recommence cet entraînement.

■ ÉTABLI

Dans les documents cités, la mise à jour des poids passe par des exemples et par un travail d'entraînement, ou par la procédure décrite pour InstructGPT. Elle n'est pas présentée comme l'effet immédiat d'une question isolée.

Une architecture n'est pas tous les systèmes

Le résumé de Vaswani et de ses collègues, en 2017, décrit le Transformer. L'architecture proposée s'appuie sur des mécanismes d'attention. Elle se passe, dans cette proposition, de récurrence et de convolution [SRC-002]. Les résultats rapportés concernent d'abord la traduction. Le résumé ne décrit pas, à lui seul, les assistants conversationnels disponibles en 2026.

Il est donc utile, et insuffisant. Utile, parce qu'il montre qu'une représentation calculée peut naître d'une architecture précise, et non d'une bibliothèque de phrases stockées telles quelles. Insuffisant, parce que les systèmes ultérieurs ajoutent des données, des objectifs, des adaptations et des outils que cet article ne catalogue pas.

■ INCONNU

Le résumé de 2017 n'autorise pas à décrire le fonctionnement interne de chaque modèle commercial actuel. Cette description resterait à établir produit par produit.

Ce que l'erreur permet de dire

On peut expliquer l'entraînement sans équation. La chaîne pédagogique est :

exemples sélectionnés → objectif d'apprentissage → calcul d'erreur → ajustement des paramètres → évaluation sur des exemples distincts.

Le système produit une réponse provisoire. On la compare à un objectif fixé à l'avance : la suite attendue, l'étiquette, la préférence d'un évaluateur. L'écart sert à modifier les poids. Cette présentation est une image de travail. Elle ne remplace pas la lecture d'un article de méthode, et aucun des résumés mobilisés ici n'a été traité comme une démonstration ligne à ligne.

Exemple pédagogique : classer un genre documentaire

Reprenons la tâche du chapitre 6, sans prétendre l'avoir exécutée. On constitue des exemples de notes, de lettres et de textes publics. L'objectif est d'estimer un genre. Après ajustement, on évalue sur des pièces tenues à l'écart. Si les exemples ne contiennent presque que des lettres, le système pourra bien réussir sur les lettres et échouer sur les notes. Le résultat dépend des exemples, de l'objectif et de l'évaluation - pas du seul volume.

L'image suffit pourtant à éviter une confusion. Un réseau n'est pas « une base de données qui aurait tout compris ». Il est le résultat d'un objectif, d'exemples sélectionnés et d'ajustements. Changer l'objectif, ou changer les exemples, peut changer ce que les poids rendent probable. Le volume brut ne garantit pas la qualité. Aucun des résumés mobilisés ici ne dit que plus d'exemples produisent, à eux seuls, un meilleur modèle.

Le chapitre suivant demande ce que l'on appelle un modèle de fondation lorsque le même socle est réutilisé pour des tâches très différentes.

À RETENIR

- 01 Exemples, objectif, erreur, paramètres et évaluation forment une chaîne.
 - 02 Les exemples d'entraînement et les poids obtenus ne sont pas le même objet.
 - 03 Une tâche de classification documentaire montre pourquoi le résultat dépend du corpus choisi.
-

04 Le Transformer de 2017 n'est pas le portrait de tous les assistants de 2026.

Sources mobilisées

- **SRC-002** - Vaswani et al., 2017. Notice et résumé. Architecture d'attention. Pas une description des modèles conversationnels ultérieurs.
- **SRC-010** - Ouyang et al., 2022. Notice et résumé. Ajustement puis préférences pour InstructGPT. Portée limitée à l'étude.
- **SRC-039** - Ajustement supervisé OpenAI. Page consultée le 9 octobre 2026. Disponibilité de la plateforme en évolution.

9 · Les modèles de fondation

Un socle entraîné largement, puis adapté, n'est ni une intelligence générale ni un simple fichier de textes.

Un nom pour une pratique

En 2021, Bommasani et ses collègues proposent d'appeler modèles de fondation des modèles entraînés sur des données larges, puis adaptés à plusieurs tâches. Leur rapport en discute les possibilités et les risques, parmi lesquels une homogénéisation : les mêmes défauts peuvent se retrouver dans de nombreuses applications construites sur un socle commun [SRC-003]. Le document est une synthèse datée. Les extraits consultés ne sont pas le rapport entier, et 2021 n'est pas 2026.

Le nom est utile s'il reste descriptif. Il désigne une manière de construire et de réutiliser un système. Il ne prouve pas qu'un modèle « comprend tout », ni que deux produits partageant ce nom ont le même corpus, le même objectif ou les mêmes limites.

■ ÉTABLI

Le rapport de 2021 définit une catégorie et en signale des risques, dont l'homogénéisation. Il ne mesure pas les produits commercialisés cinq ans plus tard.

Deux façons de s'adapter

Une adaptation peut modifier les poids. C'est le sens de l'ajustement décrit pour InstructGPT : des démonstrations, puis des préférences humaines, sur une famille de modèles [SRC-010]. C'est aussi le sens du travail d'entraînement de la documentation consultée le 9 octobre 2026 [SRC-039].

Une autre adaptation laisse les poids en place et change seulement ce qui est fourni avec la question. Le résumé de Brown et de ses collègues indique que les tâches few-shot évaluées pour GPT-3 sont réalisées sans mise à jour de gradient pendant ces tâches [SRC-009]. Des exemples placés

dans la consigne orientent la réponse présente. Ils ne constituent pas, par cela seul, un nouvel entraînement du modèle général.

Ces deux voies peuvent coexister dans un produit. Les confondre rend le débat impossible. « Le modèle a appris de moi ce matin » peut vouloir dire : la consigne a changé la réponse. Ou : un travail d’entraînement a modifié des poids. Ou encore : une mémoire de produit réinjectera plus tard une préférence. Le chapitre 11 séparera ces cas. Ici, il suffit de voir que l’adaptation n’est pas un geste unique.

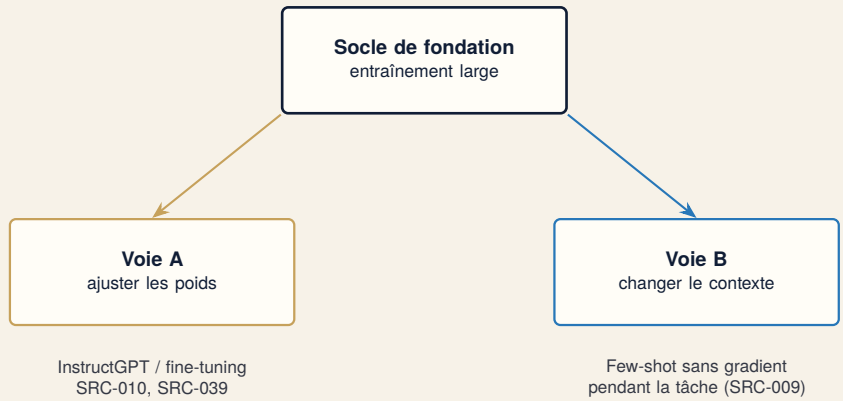


Figure. Deux adaptations distinctes d’un socle : modifier les poids, ou seulement le contexte de la réponse. Schéma explicatif. [SRC-003] [SRC-009] [SRC-010] [SRC-039].

Lecture du schéma

La voie A laisse une trace dans les paramètres. La voie B laisse une trace dans le message. Un même écran peut emprunter l’une, l’autre, ou les deux. Tant que le produit ne dit pas laquelle, la phrase « le modèle a appris » ne désigne rien d’inspectable.

Couche	Ce qu’elle apporte	Ce qu’elle ne garantit pas
Socle de fondation	Capacités générales réutilisables	Qualité uniforme sur toutes les tâches
Contexte de la question	Orientation de la réponse présente	Mise à jour des poids
Ajustement des paramètres	Adaptation sur des exemples	Couverture de tous les usages
Outils et sources externes	Accès à des documents ou services	Vérité du document retrouvé
Évaluation propre	Mesure locale d’un usage	Transfert automatique à un autre usage

Tableau. Réutiliser un socle n’équivaut pas à déployer le même système. [SRC-003] [SRC-009] [SRC-010].

Réutiliser un socle n'équivaut pas à déployer le même système. Deux outils fondés sur une famille de modèles similaire peuvent différer par leurs données, leurs adaptations, leurs outils externes et leurs critères de qualité. Les performances ne se transfèrent pas automatiquement d'un usage à un autre.

Ce que l'échelle ne règle pas

GPT-3, dans l'article de Brown et de ses collègues, est un exemple de modèle évalué à grande échelle sur des tâches peu préparées une par une. Le résumé discute aussi des difficultés et d'un risque de contamination des jeux de test [SRC-009]. Étendre ce résultat à tous les modèles, toutes les langues et toutes les questions serait une autre affirmation, que l'article ne porte pas.

Le risque signalé par Bommasani et ses collègues reste lisible sans chiffre nouveau. Si beaucoup d'outils partent du même socle, une erreur de ce socle peut se répéter partout où on l'adapte [SRC-003]. Ce n'est pas une fatalité mesurée pour 2026. C'est une raison de demander, pour chaque usage, quel socle, quelle adaptation et quelle évaluation ont été choisis.

Pour l'expérience de pensée du livre, la conséquence est directe. Un assistant qui parlerait « comme Einstein » pourrait devoir son ton à un socle général, à une adaptation sur un corpus, ou aux seuls documents placés dans la question. Tant que ces parts ne sont pas nommées, la ressemblance ne dit pas d'où elle vient.

À RETENIR

- 01 « Modèle de fondation » est une catégorie proposée en 2021, avec des risques discutés.
 - 02 Adapter peut vouloir dire changer les poids ou seulement le contexte.
 - 03 Socle, contexte, ajustement, outils et évaluation forment des couches distinctes.
 - 04 L'échelle d'un résultat ne le transforme pas en loi de tous les systèmes.
-

Sources mobilisées

- SRC-003 - Bommasani et al., 2021. Résumé et extraits. Définition et risques, dont l'homogénéisation. Rapport non lu intégralement.

- **SRC-009** - **Brown et al., 2020**. Notice et résumé. Few-shot sans mise à jour de gradient pendant les tâches décrites.
- **SRC-010** - **Ouyang et al., 2022**. Notice et résumé. Portée limitée à InstructGPT et aux consignes de l'étude.
- **SRC-039** - **Ajustement supervisé**. Page consultée le 9 octobre 2026.

10 · ChatGPT n'est pas une base de données

Une réponse fluide peut venir de poids ajustés, d'un document retrouvé, ou des deux. Ce n'est pas la même opération.

Ce que ce chapitre ajoute

Le chapitre précédent a situé les modèles de fondation. Celui-ci isole une seule confusion : prendre une phrase fluide pour une fiche retrouvée. Le chapitre suivant fixera le vocabulaire stable. Celui d'après passera au consentement et aux destinations d'une conversation.

Retrouver n'est pas continuer

Une base de données, dans l'usage ordinaire du mot, conserve des enregistrements et peut en rendre un lorsqu'on l'interroge. On peut se tromper de clé. On peut recevoir une fiche incomplète. On ne reçoit pas, en principe, une phrase inventée pour combler le vide.

Un modèle de langage fait autre chose lorsqu'il poursuit un texte à partir de ses paramètres. La suite produite peut ressembler à un souvenir exact. Elle peut aussi être seulement probable au regard de l'entraînement. La fluidité ne dit pas lequel des deux cas on a sous les yeux.

Le résumé de Lewis et de ses collègues pose une distinction qui sert encore. Une mémoire paramétrique est ce que les poids encodent. Une mémoire non paramétrique est un ensemble de documents que le système peut retrouver et fournir au modèle au moment de répondre [SRC-004]. Leur article décrit une architecture expérimentale précise. Il ne décrit pas tous les produits qui, depuis, disent « faire de la recherche ».

■ ÉTABLI

Le résumé de Lewis et al. sépare connaissances portées par les paramètres et documents retrouvés à la demande. Cette séparation ne décrit pas, à elle seule, chaque produit commercial.

Trois phrases qui ne disent pas la même chose

« Le modèle le sait. » On affirme alors que les poids rendent cette réponse probable, sans document ajouté pour l'occasion.

« Le système l'a retrouvé. » On affirme qu'un passage externe a été cherché, puis placé dans le contexte de la réponse.

« On le lui a montré dans la question. » On affirme que l'exemple fait partie du message en cours. Dans les tâches few-shot décrites pour GPT-3, ce geste change la réponse sans mise à jour de gradient pendant l'évaluation [SRC-009].

Un écran unique peut réaliser l'une de ces trois opérations, ou les combiner, sans que la phrase finale porte l'étiquette. Demander la provenance n'est pas un raffinement d'expert. C'est la seule façon de savoir ce qui a été utilisé.

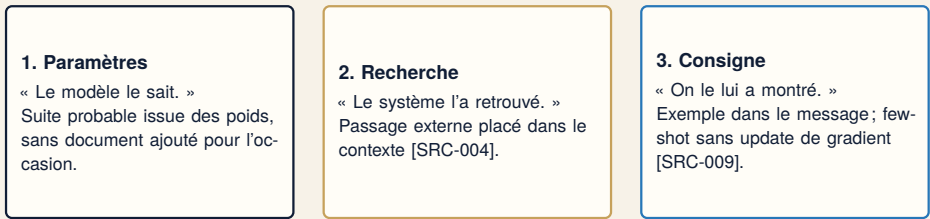


Figure. Trois provenances possibles d'une réponse fluide. Sans étiquette, l'écran unique les confond. Schéma conceptuel. [SRC-004] [SRC-009].

Situation	Ce que le système fait	Vérification nécessaire
Recherche documentaire	Retrouve un passage	Document, page, contexte
Génération paramétrique	Produit une suite probable	Chaque affirmation importante
Génération avec recherche	Combine documents et génération	Fidélité de la réponse aux sources

Tableau. Trois opérations distinctes derrière une même fenêtre. La recherche n'est pas une garantie de vérité. [SRC-004] [SRC-009].

La même question, trois opérations

Reprenons la question du chapitre 4 sur l'évolution d'une formulation entre deux versions. Trois traitements possibles :

1. **Retrouver** les deux passages dans le corpus et les comparer. 2. **Générer** une réponse à partir des seules régularités des paramètres. 3. **Générer après recherche** : placer des passages retrouvés dans le contexte, puis produire une synthèse.

La vérification n'est pas la même. Dans le premier cas, on contrôle le document. Dans le second, chaque affirmation. Dans le troisième, la fidélité

de la synthèse aux sources. La recherche documentaire n'est jamais une garantie de vérité : une source incorrecte, hors contexte ou mal lue peut aussi conduire à l'erreur.

Pourquoi la distinction compte pour Einstein

Si l'assistant « cite » une lettre de 1905, il faut savoir si un volume a été retrouvé, si un extrait a été collé dans la question, ou si une suite probable a été inventée à partir des poids. Les trois cas ne méritent pas le même crédit historique.

Ce qu'une réponse convaincante ne prouve pas

Une réponse peut être juste parce que le document retrouvé l'était. Elle peut être fausse parce que le document l'était, ou parce qu'aucune source n'a été consultée et que la suite probable a comblé un trou. Elle peut être juste par coïncidence. Le registre de ce livre ne contient pas, à ce stade, une mesure du taux d'erreur de ChatGPT en 2026. Il serait malhonnête d'en inventer un.

La conséquence pour Einstein est plus nette que n'importe quel pourcentage. Si l'assistant cite une lettre, il faut pouvoir retrouver la lettre dans une édition identifiée. Si l'assistant seulement « sonne juste », on a une imitation de surface. Les chapitres précédents ont déjà refusé de prendre cette imitation pour une identité. Ce chapitre ajoute le mécanisme : sans distinguer paramètres et documents retrouvés, on ne sait même pas quelle sorte d'imitation on juge.

À RETENIR

- 01 Recherche, génération et génération assistée par recherche sont trois opérations.
- 02 Continuer un texte à partir des poids n'est pas rendre une fiche conservée.
- 03 La recherche n'est pas une garantie de vérité.
- 04 Aucune phrase fluide ne porte, à elle seule, sa provenance.

Sources mobilisées

- **SRC-004** - Lewis et al., 2020. Notice et résumé. Distinction parametric / non-parametric. Pas tous les produits ultérieurs.

- **SRC-009 - Brown et al., 2020.** Notice et résumé. Pas de mise à jour de gradient pendant les tâches few-shot décrites.

11 · Le contexte, la mémoire et les paramètres

Quatre mots doivent rester stables pour la suite du livre : paramètres, contexte, mémoire de produit, recherche.

Le vocabulaire qui manquait

Les chapitres 8 à 10 ont montré des mécanismes. Celui-ci ne les redémontre pas : il les nomme une fois pour toutes. Les parties suivantes devront utiliser ces quatre cases, ou recommencer la confusion.

Quatre termes suffisent, à condition de ne pas les échanger.

Les **paramètres** sont les poids issus d'un entraînement antérieur. Les modifier demande un travail sur des exemples, tel que le décrit la documentation d'ajustement supervisé consultée le 9 octobre 2026 [\[SRC-039\]](#).

Le **contexte** est ce qui est fourni pour la réponse en cours : la question, les consignes, les exemples collés dans le message, les passages qu'un outil vient d'ajouter.

Une **mémoire de produit** conserve certaines informations pour les réinjecter plus tard, selon les réglages du service. La documentation de ChatGPT consultée le même jour sépare ces contrôles de mémoire des contrôles qui portent sur l'usage de conversations pour améliorer les modèles [\[SRC-006\]](#) [\[SRC-007\]](#).

Une **recherche externe** retrouve des documents et les place dans le contexte. Elle ne réécrit pas, par ce seul geste, les poids [\[SRC-004\]](#).

■ ÉTABLI

Dans les pages consultées le 9 octobre 2026, mémoriser une préférence, retrouver un document et ajuster des poids sont trois opérations distinctes. Cette séparation décrit un produit et une date, pas une loi de tous les systèmes.

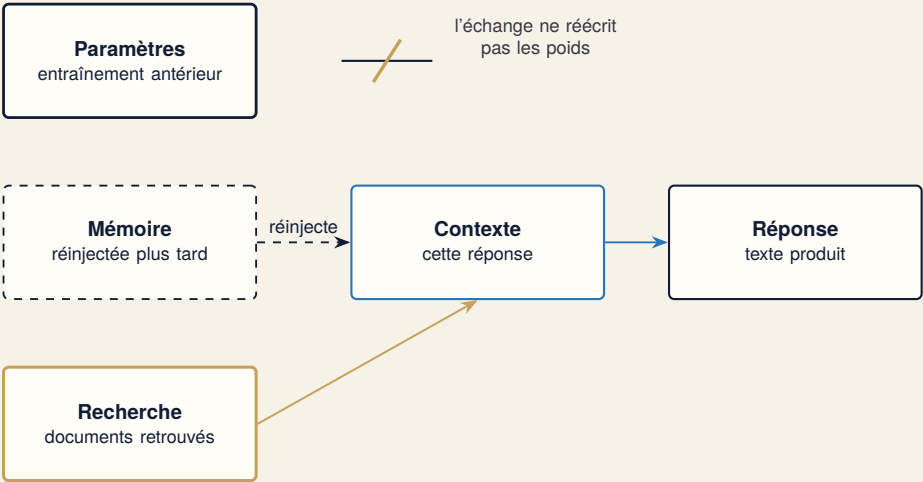


Figure. Paramètres, contexte, mémoire de produit et recherche externe n'entrent pas dans le même geste. L'échange alimente le contexte. Il ne met pas à jour les poids. Schéma explicatif, non un résultat expérimental. [SRC-006] [SRC-007] [SRC-009] [SRC-039], aux niveaux d'examen du registre, documentation produit datée du 9 octobre 2026.

Mécanisme	Ce qui change	Ce qui ne change pas
Paramètres	Les poids après un travail d'entraînement	Une question isolée dans la fenêtre
Contexte	La réponse en cours (message, outils)	Les poids du modèle général
Mémoire de produit	Ce qui peut être réinjecté plus tard	L'autobiographie d'une personne
Recherche externe	Les documents placés dans le contexte	Les paramètres, par ce seul geste

Tableau. Quatre mécanismes à ne pas échanger. Descriptions datées pour la documentation produit consultée le 9 octobre 2026. [SRC-004] [SRC-006] [SRC-007] [SRC-039].

Ce qu'une conversation peut faire, et ne pas faire

Une phrase envoyée dans une fenêtre entre dans un contexte. Elle peut changer la réponse suivante. Dans les tâches few-shot évaluées par Brown et ses collègues, ce changement a lieu sans mise à jour de gradient pendant ces tâches [SRC-009].

La même phrase ne devient pas, par son envoi, un nouvel entraînement du modèle général. La documentation des contrôles de données distingue l'historique et les choix relatifs à l'amélioration des modèles [SRC-006]. Une politique peut prévoir un usage ultérieur. Cet usage aurait lieu dans une autre temporalité, avec d'autres contrôles. Il ne se confond pas avec la réponse que l'on vient de lire.

La conversation temporaire, dans la page consultée le 9 octobre 2026, est présentée comme un mode sans mémoire de produit et sans usage pour l'entraînement déclaré. La page signale aussi des rétentions de sécurité et des exceptions [SRC-008]. « Temporaire » n'est donc pas un synonyme de « rien n'est conservé, en aucun sens ».

■ ÉTABLI

Une conversation ne doit pas être décrite comme une modification instantanée des paramètres du modèle général. Les exceptions de rétention et les politiques ultérieures doivent rester nommées, et datées.

Diagnostic : « l'assistant semble se souvenir »

Quand une réponse donne l'impression d'un souvenir, quatre explications doivent être examinées avant toute conclusion :

1. **Contexte encore accessible** - l'information figure dans le message ou dans l'historique de la conversation en cours. 2. **Mémoire de produit** - une préférence ou un fait a été enregistré et réinjecté selon les réglages du service. 3. **Recherche externe** - un document a été retrouvé et placé dans le contexte. 4. **Régularité paramétrique** - les poids rendent cette suite probable sans document ajouté pour l'occasion.

L'apparence de la réponse ne décide pas seule du mécanisme. Cette séquence de diagnostic est la carte de référence du livre. Les chapitres suivants devront y revenir plutôt que de redéfinir les mêmes mots.

Pourquoi le livre s'arrête sur ces quatre mots

Tout le scénario d'Einstein dépend de cette grille. Un corpus historique placé dans une recherche n'est pas une mémoire autobiographique. Une préférence de style gardée par le produit n'est pas un carnet de 1905. Un ajustement des poids sur des lettres éditées ne serait pas une conversation avec l'auteur de ces lettres. Et aucun de ces gestes ne rendrait présent un homme mort en 1955.

Les parties suivantes pourront parler de traces humaines, de modèles du monde ou de boucles entre textes générés et nouveaux entraînements. Elles le feront avec ces quatre cases, ou elles recommenceront la confusion que cette partie vient de défaire.

La vérification, elle, n'est pas close. Les pages de produit changent. Les résumés d'articles ne sont pas les articles. La note placée avant le prologue le dit pour l'ensemble du livre : une réponse convaincante ne tient pas lieu de preuve. Ici, elle s'applique au vocabulaire lui-même. Si une documen-

tation évolue, la case correspondante devra être réécrite, pas défendue par habitude.

À RETENIR

- 01 Paramètres, contexte, mémoire de produit et recherche externe forment la carte de référence.
- 02 Quand un système « semble se souvenir », quatre explications doivent être examinées.
- 03 Une question change d'abord un contexte ; elle ne met pas à jour les poids par elle-même.
- 04 Cette grille sera réutilisée, non redéfinie, dans la suite du livre.

Sources mobilisées

- **SRC-004 - Lewis et al.** Notice et résumé. Récupération distincte de la mémoire paramétrique.
- **SRC-006 - Contrôles de données OpenAI.** Page consultée le 9 octobre 2026.
- **SRC-007 - Mémoire OpenAI.** Page consultée le 9 octobre 2026. Fonctions variables selon l'offre.
- **SRC-008 - Conversation temporaire.** Page consultée le 9 octobre 2026. Rétentions de sécurité à ne pas oublier.
- **SRC-009 - Brown et al.** Notice et résumé. Pas de mise à jour de gradient pendant les tâches few-shot décrites.
- **SRC-039 - Ajustement supervisé.** Page consultée le 9 octobre 2026.

12 · Chaque conversation est-elle une donnée ?

Une phrase envoyée existe. Elle ne devient une donnée d'entraînement que si une chaîne de traitement le décide.

La question

Le chapitre 11 a fixé quatre mécanismes. Ici, la question change : non plus « comment la réponse est calculée », mais **où va la trace une fois envoyée** ? C'est un problème de gouvernance et de consentement.

Cas fictif signalé

■ EXPÉRIENCE DE PENSÉE

Le cas suivant est fictif. Il sert à rendre visibles des destinations possibles, non à décrire un usage réel d'un utilisateur nommé.

Un utilisateur demande une synthèse. Il corrige la réponse. Plus tard, il retrouve une préférence de format. Trois événements distincts ont pu se produire : le message a influencé le contexte ; une préférence a été mémorisée ; rien n'indique encore qu'un pipeline d'entraînement a utilisé cet échange. Confondre les trois, c'est croire que « chaque conversation entraîne immédiatement le modèle ».

Cycle de vie conditionnel

Entre la saisie et une éventuelle réutilisation, plusieurs étapes peuvent intervenir : traitement de la réponse, conservation, contrôle d'accès, filtrage, sélection, préparation, puis - seulement si un pipeline approprié existe - entraînement ou évaluation.



Branches conditionnelles selon produit, réglages et politiques. Pas une loi de tous les services.

Figure. Cycle de vie conditionnel d'une conversation : message, contexte, destinations, sélection, éventuel pipeline. Schéma conceptuel ; dépend du produit et des politiques. [SRC-006] [SRC-007] [SRC-008].

Fonction	Question	Ce qu'elle ne prouve pas
Contexte actif	Influence la réponse actuelle ?	Apprentissage durable des poids
Mémoire de produit	Réutilisation ultérieure possible ?	Changement des paramètres généraux
Recherche documentaire	Contenu retrouvé à la demande ?	Vérité du contenu
Entraînement	Exemples utilisés dans un pipeline ?	Intégration de chaque conversation

Tableau. Une trace, quatre fonctions. Distinguer techniquement possible, effectivement mis en œuvre et autorisé par une politique.

La documentation de ChatGPT consultée le 9 octobre 2026 distingue les contrôles de mémoire et ceux qui portent sur l'usage de conversations pour améliorer les modèles [SRC-006] [SRC-007]. Le mode temporaire est présenté sans mémoire de produit et sans usage pour l'entraînement déclaré, avec des rétentions de sécurité [SRC-008]. Ces pages décrivent un produit et une date, non une loi de tous les systèmes.

■ ÉTABLI

Dans les pages consultées le 9 octobre 2026, mémoire de produit et usage pour l'amélioration des modèles sont des contrôles séparés.

Possible, mis en œuvre, autorisé

Il faut encore distinguer trois niveaux. **Techniquement possible** : une architecture pourrait conserver ou réutiliser une trace. **Effectivement mis en œuvre** : un service le fait, pour certaines données, à certaines dates. **Autorisé par une politique** : les réglages, le consentement, la durée de conservation et la suppression le permettent. Sans ces trois niveaux, le mot « donnée » efface la décision.

Consentement, confidentialité, durée, suppression et traçabilité ne sont pas des annexes. Ils décident si une conversation reste un contexte, devient une mémoire de produit, ou entre - éventuellement - dans un pipeline.

Conséquence

Les échanges imaginaires d'Einstein avec un assistant ne seraient pas automatiquement un corpus d'apprentissage. Sans fiche de traitement, la scène ne dit rien de ce que le système apprendrait. Le lecteur doit pouvoir réfuter l'affirmation simpliste d'un entraînement immédiat à chaque message.

À RETENIR

- 01 Une conversation laisse des traces ; elle n'entraîne pas automatiquement.
 - 02 Contexte, mémoire, recherche et entraînement sont quatre fonctions.
 - 03 Possible, mis en œuvre et autorisé ne se confondent pas.
 - 04 Sans destination nommée, le mot « donnée » reste trop large.
-

Sources mobilisées

- **SRC-006 - Contrôles de données.** Page consultée le 9 octobre 2026.
- **SRC-007 - Mémoire.** Page consultée le 9 octobre 2026.
- **SRC-008 - Conversation temporaire.** Page consultée le 9 octobre 2026.
- **SRC-009 - Brown et al.** Notice et résumé. Few-shot sans mise à jour de gradient pendant les tâches décrites.

13 · Le chercheur nourrit-il l'IA ?

Publications, démonstrations et préférences humaines peuvent entrer dans une chaîne d'apprentissage. Elles n'y entrent pas toutes de la même façon.

La question

Le chercheur nourrit-il l'IA ? La formulation suggère un destin. La vraie question est : **par quelles médiations une production scientifique peut-elle modifier les matériaux, les critères ou les comportements d'un système ?**

Cas en trois temps

■ **EXPÉRIENCE DE PENSÉE**

Chaîne conditionnelle illustrée. Aucune entrée automatique dans un entraînement n'est affirmée.

1. Une étude est publiée avec méthode, résultats et limites. 2. Une IA en produit une synthèse claire. 3. Cette synthèse est reprise dans un cours ou un article secondaire.

Chaque médiation peut préserver, simplifier, déformer ou perdre une information. Une sortie générée n'entre pas automatiquement dans un futur entraînement. Elle peut toutefois devenir une nouvelle trace publique - et, si un pipeline la collecte, une donnée candidate.



Figure. De la publication à une éventuelle réintégration. Chaque médiation peut préserver, simplifier ou déformer. Schéma conditionnel de l'auteur.

Élément	Risque	Contrôle
Résultat	Généralisation abusive	Retour à l'article original
Méthode	Omission d'une étape	Vérification du protocole
Limite	Disparition dans le résumé	Conserver les réserves
Erreur	Amplification	Provenance et validation
Incertitude	Devenue certitude	Niveaux de confiance

Tableau. Ce qui se transmet d’une publication à une synthèse. Une sortie générée n’entre pas automatiquement dans un entraînement.

Articles, résumés, commentaires, sorties

Il faut séparer les statuts. L’article original porte la méthode et les réserves. Le résumé peut les écarter. Le commentaire ajoute une interprétation. La sortie générée peut transformer une incertitude en certitude. Messeri et Crockett analysent le risque d’illusions de compréhension en science [SRC-013]. Un outil qui accélère une tâche peut faire croire qu’il a saisi le domaine.

Le résumé d’Ouyang et de ses collègues décrit, pour InstructGPT, des démonstrations et des préférences humaines comme signaux d’ajustement [SRC-010]. Ce n’est ni la lecture d’un article par un humain, ni une conversation ordinaire.

■ ÉTABLI

Le résumé d’InstructGPT décrit des démonstrations et des préférences humaines comme sources d’ajustement. Il ne décrit pas tous les usages scientifiques de l’IA.

Pourquoi la clarté peut tromper

Un chercheur peut préférer une réponse parce qu’elle est claire. La clarté n’est pas la correction. Si l’erreur est ensuite reprise sans retour à la source, la chaîne amplifie une apparence de savoir. Le contrôle minimal reste le retour à l’article original et la conservation des limites.

Conséquence

Nourrir un modèle de textes scientifiques, ce n’est pas lui transmettre le métier entier. C’est lui donner des traces sélectionnées - et parfois des médiations déjà déformées. Le chapitre suivant demande ce que ces traces laissent voir des méthodes.

À RETENIR

- 01 Publier rend une trace disponible ; cela ne décide pas seul de son usage.
 - 02 Chaque médiation peut perdre une limite ou amplifier une erreur.
 - 03 Clarté et correction ne se confondent pas.
 - 04 Une sortie générée n'entre pas automatiquement dans un entraînement.
-

Sources mobilisées

- **SRC-001** - **Einstein Papers Project**. Portée d'un corpus édité, non exhaustivité mentale.
- **SRC-010** - **Ouyang et al.** Notice et résumé. Démonstrations et préférences.
- **SRC-013** - **Messeri et Crockett**. Notice et résumé. Illusions de compréhension en science.

14 · L'IA apprend-elle aussi nos méthodes ?

Des textes peuvent montrer des procédures. Ils ne montrent pas toujours pourquoi on les a choisies, ni ce qu'on a écarté.

La méthode dans le texte

Une démonstration, une preuve écrite, un protocole ou un code commenté rendent visible une manière de faire. Un modèle entraîné sur de tels textes peut reproduire certaines régularités : l'ordre des étapes, le vocabulaire d'une discipline, la forme d'une démonstration. Le résumé d'InstructGPT montre qu'un système peut être ajusté pour mieux suivre des instructions formulées par des humains [SRC-010]. Suivre une instruction n'est pas encore maîtriser une méthode scientifique.

Norton rappelle, dans l'introduction lue pour ce livre, que des récits populaires de découverte simplifient parfois des enquêtes longues [SRC-040]. Si un corpus contient surtout ces récits, un modèle peut apprendre la simplification plutôt que la méthode. La ressemblance de style n'est pas la fidélité au travail.

Étude de cas documentaire - trois questions

Reprenons la grille du chapitre 1 sur un dossier historique vérifié en catégories, non en intentions inventées. Les notices des *Collected Papers* séparent déjà genres et périodes [SRC-033] [SRC-035] [SRC-036].

1. **Observables.** Quelles modifications de formulation sont directement visibles entre versions datées ? 2. **Inférences prudentes.** Quels ordres d'opérations ou types de corrections peut-on comparer sans prétendre lire l'intention ? 3. **Inconnus.** Quelles motivations, hésitations non écrites ou contextes de laboratoire restent hors pièces ?

Ne pas attribuer à Einstein une intention psychologique que les pièces ne permettent pas d'établir. Apprendre des connaissances produites et ap-

prendre des régularités de méthode sont deux objets distincts.

Ce que les démonstrations omettent

Une démonstration réussie montre un chemin. Elle montre moins les erreurs non publiées, les instruments indisponibles, les négociations d'équipe ou les normes locales. Premack et Woodruff définissaient déjà l'attribution d'états mentaux comme une inférence sur ce qui n'est pas observable directement [SRC-027]. Appliquée aux méthodes, la même prudence vaut : on infère une procédure à partir de traces. On n'observe pas l'intention entière.

■ INCONNU

Aucun des résumés mobilisés ici ne mesure la part de méthode scientifique réellement capturée par un modèle de langage contemporain.

Conséquence pour Einstein

Si l'on voulait qu'un assistant « apprenne la méthode d'Einstein », il faudrait d'abord préciser laquelle : rédaction d'un article, correction d'une note, échange épistolaire, ou récit ultérieur d'une découverte. Les notices des Collected Papers séparent déjà des genres de documents [SRC-033] [SRC-035] [SRC-036]. Les confondre produirait une méthode moyenne qui n'aurait existé nulle part.

L'acquis de cette partie est donc simple. Une trace humaine peut devenir une donnée. Elle le devient par des choix. Et ce qu'elle enseigne dépend de ce qu'elle montre, de ce qu'elle cache, et de l'objectif d'apprentissage.

À RETENIR

- 01 Des textes peuvent encoder des régularités de procédure sans transmettre le métier entier.
 - 02 Les récits simplifiés peuvent être plus faciles à imiter que les enquêtes longues.
 - 03 Attribuer une méthode reste une inférence à partir de traces.
 - 04 Les genres documentaires ne sont pas interchangeables.
-

Sources mobilisées

- SRC-010 - Ouyang et al. Notice et résumé.

- **SRC-027 - Premack et Woodruff.** Notice et résumé. Inférence sur des états non observés.
- **SRC-033, SRC-035, SRC-036 - Notices de volumes.** Genres documentaires, pas pièces individuelles.
- **SRC-040 - Norton.** Introduction et cadrage. Mythes de découverte.

15 · Qu'est-ce qu'un « modèle du monde » ?

Le même mot désigne des objets différents en sciences cognitives, en apprentissage par renforcement et dans le langage courant de l'IA.

Un mot trop chargé

« Modèle du monde » suggère une petite copie intérieure de la réalité. L'image est forte. Elle est aussi dangereuse. Une représentation peut servir à prédire ou à agir sans ressembler au monde comme une carte ressemble à un paysage.

La présentation éditoriale de Johnson-Laird décrit des modèles mentaux comme des structures qui permettent de raisonner sur des relations entre objets et événements [SRC-018]. Cette formulation vient d'une notice d'éditeur. L'ouvrage intégral n'a pas été lu pour ce livre. Elle suffit à rappeler qu'en sciences cognitives, « modèle » n'est pas synonyme de photographie intérieure.

Un usage précis en apprentissage

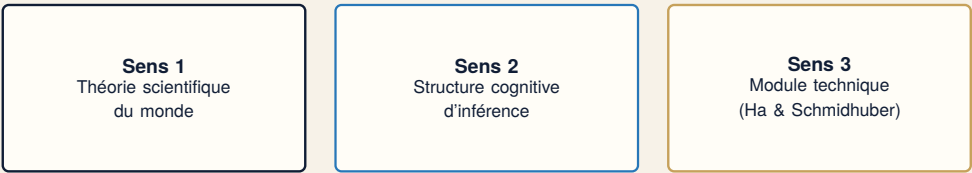
Ha et Schmidhuber appellent « world models » des modèles génératifs d'environnements, utilisés notamment pour apprendre dans des cadres d'apprentissage par renforcement [SRC-011]. Le résumé décrit des représentations spatio-temporelles compressées. Il ne permet pas de conclure que tout modèle de langage possède, au même sens, un modèle du monde.

■ ÉTABLI

Dans le résumé de Ha et Schmidhuber, « world model » désigne un cadre précis d'apprentissage. Ce cadre ne s'étend pas automatiquement à tous les LLM.

Trois sens à ne pas fusionner

Le premier sens est descriptif : une théorie scientifique du monde physique ou social. Le deuxième est cognitif : une structure interne qui guide l'inférence ou l'action. Le troisième est technique : un module qui prédit des états d'un environnement pour un agent. Un assistant conversationnel peut exceller à produire des phrases sur le monde sans que l'on ait montré lequel de ces trois sens s'applique.



Trois objets distincts. Les fusionner transforme une métaphore en preuve.

Figure. Trois sens de « modèle du monde ». Le livre exige un complément avant d'employer l'expression. Schéma conceptuel. [SRC-011] [SRC-018].

Sens	Objet	Question pertinente
Théorie scientifique	Explication d'un phénomène	Quelles observations la soutiennent ?
Modèle cognitif	Structure pour raisonner	Quelles inférences permet-elle ?
Modèle technique	Prédiction d'états	Quelles prédictions sont correctes ?
Représentation LLM	Structures apprises débattues	Quels tests les caractérisent ?

Tableau. Quatre sens du mot « modèle ». Non équivalents. [SRC-011] [SRC-018].

Critère d'emploi dans le livre

Pour la suite, nous n'emploierons « modèle du monde » qu'avec un complément : modèle du monde au sens de Ha et Schmidhuber, représentation utile pour une tâche, ou discours populaire. Sans ce complément, le chapitre suivant ne pourrait pas poser correctement la question de savoir ce qu'un système possède réellement.

Un test simple aide le lecteur. Si l'on peut remplacer la formule par « prédiction d'états pour un agent », on est dans le sens technique. Si l'on veut dire « théorie scientifique », il faut le dire. Si l'on veut dire « structure mentale humaine », il faut une source cognitive, non une démonstration de chatbot.

À RETENIR

- 01 « Modèle du monde » a plusieurs sens disciplinaires.
 - 02 Johnson-Laird est cité ici via une notice d'éditeur, non via une lecture intégrale.
 - 03 Ha et Schmidhuber définissent un usage technique limité.
 - 04 Fusionner les sens transforme une métaphore en preuve.
-

Sources mobilisées

- **SRC-011** - Ha et Schmidhuber. Notice et résumé.
- **SRC-018** - Johnson-Laird. Présentation éditoriale seulement.

16 · L'IA possède-t-elle une représentation du monde ?

Des performances et des débats existent. Un verdict unique sur la compréhension n'est pas établi par les sources de ce livre.

Ce que l'on peut mesurer

On peut mesurer ce qu'un système prédit, classe ou génère. On peut comparer ces sorties à des références. On peut parfois inspecter des états internes. Aucune de ces opérations, à elle seule, ne décide si le système « possède une représentation du monde » au sens où un humain en parle.

Mitchell et Krakauer cartographient le débat sur la compréhension des grands modèles de langage et insistent sur la nécessité de critères différenciés [SRC-019]. Leur article de perspective ne fournit pas un verdict expérimental unique. Bender et Koller soutiennent, dans un article de position, qu'un entraînement sur la seule forme linguistique ne suffit pas, à lui seul, à établir l'apprentissage de la signification [SRC-020]. C'est un argument d'auteurs, non une mesure de tous les systèmes multimodaux ultérieurs.

■ INCONNU

Les sources consultées n'autorisent pas à conclure que les modèles de langage contemporains possèdent, ou non, une représentation du monde au sens humain.

Cinq niveaux d'évaluation

Avant de parler de représentation du monde, il faut savoir quel test a été mené.

Niveau	Question	Conclusion autorisée
Performance	Réussite d'une tâche ?	Réussite dans ces conditions
Généralisation	Cas nouveaux ?	Généralisation dans le périmètre
Robustesse	Reformulations ?	Stabilité relative
Relations	Liens causaux ou temporels ?	Certaines relations tenues ou non
Mécanismes	Indices internes testables ?	Hypothèses sur les représentations

Tableau. Cinq niveaux d'évaluation. Aucun ne mesure à lui seul la compréhension humaine. [SRC-019] [SRC-020].

Aucun de ces niveaux ne constitue, à lui seul, une mesure de la compréhension humaine ou de la conscience. Le lecteur doit refuser une conclusion plus forte que le test ne le justifie.

Ce que l'on ne doit pas confondre

Ha et Schmidhuber décrivent des world models dans un cadre d'apprentissage par renforcement [SRC-011]. Ce cadre n'est pas ChatGPT. Shanahan rappelle, dans un article de synthèse, la nécessité d'un langage qui reflète les mécanismes des modèles et évite des imputations injustifiées [SRC-025]. Attribuer une compréhension globale parce qu'une réponse paraît informée, c'est précisément l'imputation à éviter.

Une représentation utile pour une tâche peut exister sans expérience subjective, sans intention et sans carte complète du monde. Le livre l'a déjà dit pour une personne. Il le répète pour un système.

Conséquence pour l'enquête

Si l'assistant d'Einstein répond correctement à une question de physique, on a une performance. Si la réponse cite un document retrouvé, on a une provenance. Si les poids rendent la suite probable, on a une régularité encodée. Aucun de ces cas ne tranche, à lui seul, le débat sur la compréhension.

À RETENIR

- 01 Mesurer une performance n'équivaut pas à trancher la compréhension.
- 02 Mitchell et Krakauer cartographient un débat, sans verdict unique.
- 03 Bender et Koller formulent un argument de position sur forme et signification.
- 04 Le vocabulaire doit coller au mécanisme, non à l'impression.

Sources mobilisées

- **SRC-011** - Ha et Schmidhuber. Notice et résumé.
- **SRC-019** - Mitchell et Krakauer. Notice et résumé.
- **SRC-020** - Bender et Koller. Résumé et extraits de conclusion.
- **SRC-025** - Shanahan. Résumé et extraits de conclusion.

17 · De millions de traces à une représentation collective

L'agrégation rend certaines régularités visibles. Elle masque aussi des absences, des langues et des choix de sélection.

La question

Que signifie « collectif » ici ? Non pas une conscience commune. Une **agrégation de productions humaines sélectionnées, transformées et pondérées** par des mécanismes techniques et institutionnels.

Cas : la répétition n'est pas la preuve

Une affirmation fréquemment reprise peut dominer un corpus. Cent documents qui répètent la même étude ne constituent pas cent confirmations indépendantes. Argyle et ses collègues montrent qu'un modèle conditionné peut reproduire certaines distributions de réponses [SRC-030]. Aher et ses collègues observent aussi une distorsion de sur-précision [SRC-031]. L'agrégation peut lisser autant qu'elle révèle.

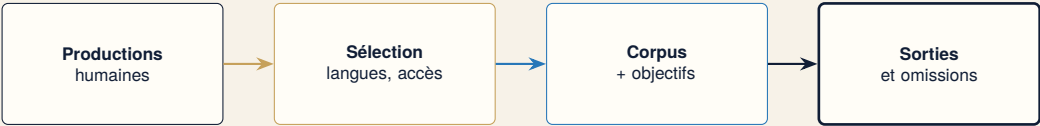
■ ÉTABLI

Dans les études citées, l'agrégation peut reproduire certaines distributions collectives. Elle ne démontre pas la restitution d'une personne.

Carte des absences

Avant de célébrer le volume, dresser les absences possibles : langues, périodes, régions, genres documentaires, domaines scientifiques, groupes sociaux. Bommasani et ses collègues soulignent des risques d'homogénéisation pour les modèles de fondation [SRC-003]. Une tradition orale, un savoir non publié ou une population sous-documentée n'entre pas magiquement dans la représentation.

Les archives d’Einstein rappellent le même principe à une autre échelle. Un projet éditorial rend accessibles des catégories. Il ne rend pas accessible tout ce qui a eu lieu [SRC-001] [SRC-032].



Ce qui est absent du corpus reste absent des sorties, ou y revient déformé.

Figure. Construction d’une représentation collective : productions, sélection, corpus, sorties et omissions. Schéma conceptuel. [SRC-003].

Situation	Ce qu’on pourrait croire	À vérifier
Beaucoup d’anglais	Savoir mondial couvert	Autres langues
Beaucoup de reprises	Consensus indépendant	Sources primaires
Peu de documents	Phénomène marginal	Obstacles de conservation
Beaucoup de numérisés	Archive complète	Lacunes et non-numérisés

Tableau. Abondance contre représentativité. Cent reprises d’une étude ne valent pas cent confirmations.

Diversité et indépendance

Diversité des sources et indépendance des sources ne se confondent pas. La traduction, l’indexation, la disponibilité commerciale et les droits d’accès décident aussi de ce qui devient visible. Une représentation collective utile doit être présentée comme telle : estimation sur un corpus, pour une tâche, avec des biais nommables.

Conséquence

La représentation collective n’est pas une photographie neutre de l’humanité. Elle dépend de ce que l’on a conservé, sélectionné, rendu accessible et évalué. Le chapitre suivant organise des futurs conditionnels - non une trajectoire unique.

À RETENIR

- 01 « Collectif » désigne une agrégation sélectionnée, pas une conscience.
- 02 Abondance et représentativité divergent.

03 Cent reprises ne valent pas cent preuves indépendantes.

04 Les absences d'un corpus se reproduisent dans les sorties.

Sources mobilisées

- **SRC-001, SRC-032 - Einstein Papers Project.** Portée et histoire institutionnelle.
- **SRC-003 - Bommasani et al.** Résumé et extraits.
- **SRC-030 - Argyle et al.** Notice et résumé.
- **SRC-031 - Aher et al.** Notice et résumé.

18 · 2026 → 2030 → 2040 → 2050

Des scénarios conditionnels aident à penser. Ils ne sont pas des prévisions.

■ EXPÉRIENCE DE PENSÉE

Les dates de ce chapitre ne prédisent aucun calendrier technologique. Elles organisent des conditions possibles.

Pourquoi des horizons, et non des oracles

Une expérience de pensée explore des conséquences. Elle ne transforme pas un scénario en événement [SRC-038]. Dire « en 2040, les systèmes feront X » sans conditions, c'est déguiser une préférence en chronologie.

Trois scénarios conditionnels

Scénario A - Intégration maîtrisée. Corpus mieux documentés, évaluations indépendantes, institutions de confiance. Condition : financement durable, gouvernance, accès aux sources.

Scénario B - Progrès fragmenté. Capacités en hausse, mais inégalités selon langues, domaines et droits d'accès. Condition : investissements inégalement répartis.

Scénario C - Boucles dégradées. Contenus générés réutilisés sans provenance suffisante ; erreurs répétées ; vérification en retard. Condition : vitesse de production et validation insuffisante.

A · Maîtrisé

Traçabilité, audits, sources vérifiables.
Condition : gouvernance.

B · Fragmenté

Capacités inégales selon langues et accès.
Condition : investissements.

C · Dégradé

Réutilisation sans provenance. Erreurs répétées. Condition : vitesse.

Figure. Trois scénarios conditionnels, non une chronologie inévitable. Schéma de discussion. [SRC-038].

Variable	Indicateur	Pourquoi
Qualité documentaire	Affirmations sourcées	Traçabilité
Diversité	Langues, régions, domaines	Absences visibles
Fiabilité	Erreurs sur tests indépendants	Comparaison
Gouvernance	Audit et recours	Correction possible
Dépendance	Concentration des infrastructures	Pouvoir

Tableau. Indicateurs proposés pour suivre des scénarios conditionnels. Propositions de recherche, non mesures obtenues.

Ces indicateurs sont des propositions de recherche, non des mesures déjà obtenues. Si la part des affirmations sourcées chute tandis que les contenus synthétiques dominent sans filtre, le scénario C gagne en plausibilité. Si des audits indépendants et des recours effectifs se généralisent, le scénario A devient plus crédible.

Horizon	Mécanisme plausible	Ce qui invaliderait
2026	Distinctions paramètres / contexte / mémoire / recherche	Présenter une politique produit comme loi intemporelle
2030	Intégration plus dense d'outils	Affirmer qu'une conversation égale un entraînement
2040	Boucles contenus générés corpus	Ignorer les risques d'homogénéisation déjà discutés
2050	Infrastructures de mémoire collective	Equivaloir archive collective et mémoire autobiographique

Tableau. Horizons conditionnels, non un calendrier technologique. Cadre méthodologique [SRC-038] ; documentation produit datée [SRC-006] [SRC-007].

■ **HYPOTHÈSE**

Les horizons 2030-2050 sont des cadres de discussion. Aucune des sources du registre ne fixe ces dates comme des faits.

Ce que le livre refuse

Il refuse la courbe unique du progrès. Il refuse aussi le refus total d'imaginer. Le chapitre suivant demande ce qu'une génération de modèles peut réellement hériter d'une précédente - par des pipelines, non par une biologie.

À RETENIR

01 Trois scénarios conditionnels remplacent une chronologie unique.

02 Chaque scénario a des conditions et des indicateurs.

03 Les indicateurs proposés ne sont pas des résultats obtenus.

04 Sans conditions, une date devient un récit publicitaire.

Sources mobilisées

- **SRC-003** - Bommasani et al. Risques discutés en 2021, non un calendrier.
- **SRC-006**, SRC-007 - Documentation OpenAI. Datée du 9 octobre 2026.
- **SRC-038** - Expériences de pensée. Cadre méthodologique.

19 · Chaque génération hérite-t-elle de la précédente ?

Poids, corpus, évaluations et décisions humaines se transmettent par des pipelines, pas par une généalogie naturelle.

Hériter n'est pas grandir

Une génération de modèles peut reprendre des poids, distiller un système plus grand, réutiliser un corpus, ou s'entraîner en partie sur des sorties antérieures. Chacun de ces gestes est un choix d'ingénierie. Aucun n'équivaut à une maturation biologique.

L'ajustement supervisé, dans la documentation consultée le 9 octobre 2026, part d'exemples et d'un travail qui met à jour des poids [SRC-039]. InstructGPT, dans le résumé d'Ouyang et de ses collègues, part aussi de démonstrations et de préférences [SRC-010]. Dans les deux cas, l'héritage passe par des données et des objectifs. Il ne passe pas par une mémoire vécue.

Matrice des mécanismes d'héritage

L'héritage n'est pas automatique. Il résulte de décisions d'ingénierie, de sélection et de gouvernance.

Mécanisme	Ce qui se transmet	Risque / contrôle
Poids réutilisés	Capacités antérieures	Tests indépendants
Distillation	Comportements source	Comparaison de tâches
Corpus repris	Documents et biais	Provenance
Données synthétiques	Sorties générées	Références indépendantes
Évaluations reprises	Critères et seuils	Jeux tenus à l'écart

Tableau. Héritage technique : décisions d'ingénierie, non maturation biologique.

Des compétences utiles peuvent se transmettre. Des erreurs aussi. Si des sorties synthétiques entrent dans un nouveau corpus sans filtre, une régularité trompeuse peut se renforcer. Les sources de ce livre ne mesurent pas l'ampleur de cet effet en 2026. Elles autorisent seulement à le nommer

comme risque conditionnel [SRC-003]. Toute généralisation à « tous les modèles » ou « toutes les données synthétiques » dépasserait ce qui est établi ici.

■ HYPOTHÈSE

La transmission d'erreurs via des corpus synthétiques est un risque plausible. Son intensité n'est pas établie ici par une mesure unique.

Décisions humaines dans la chaîne

Qui choisit le corpus ? Qui filtre ? Qui évalue ? Qui décide qu'un système remplace un autre ? Le cadre NIST de gestion des risques de l'IA, dans sa portée générale consultée, insiste sur des pratiques de fiabilité et de gouvernance, sans résoudre à lui seul chaque produit [SRC-005]. L'héritage technique n'efface pas la responsabilité.

Pour Einstein, la question devient : un modèle ajusté sur une édition de 2026 hériterait de cette édition, de ses choix et de ses silences, non de la vie de 1905. Le chapitre suivant examine la boucle entre humains et systèmes lorsque les sorties reviennent dans les pratiques.

À RETENIR

- 01 L'héritage entre modèles passe par des pipelines.
 - 02 Des erreurs peuvent se transmettre avec des compétences.
 - 03 Les décisions humaines restent dans la chaîne.
 - 04 Hériter d'une édition n'est pas hériter d'une vie.
-

Sources mobilisées

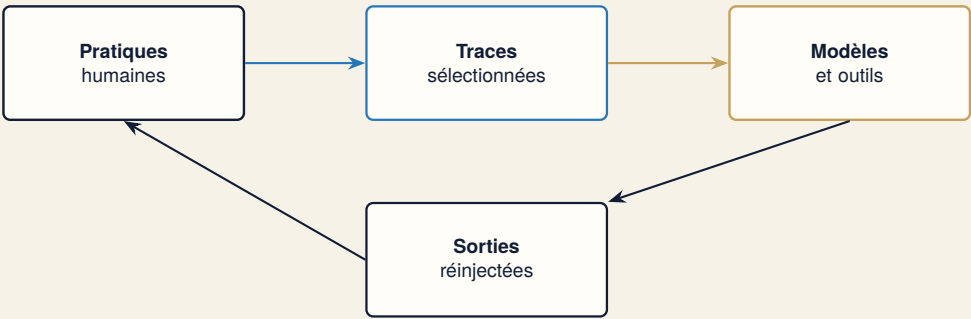
- **SRC-003** - Bommasani et al. Risques de fondation.
- **SRC-005** - NIST AI RMF. Portée générale, cadre volontaire.
- **SRC-010** - Ouyang et al. Notice et résumé.
- **SRC-039** - Ajustement supervisé. Page datée.

20 · La boucle IA ↔ humanité

Des rétroactions existent dès que des sorties entrent dans des pratiques. Elles n'améliorent pas automatiquement ce qu'elles touchent.

Une boucle n'est pas un progrès

Des humains produisent des textes. Des systèmes s'entraînent ou s'appuient sur des traces. Leurs sorties reviennent dans des articles, des codes, des conversations et des décisions. La boucle est réelle dès que cette circulation existe. Elle n'est pas, par elle-même, une garantie d'amélioration.



La boucle existe dès qu'il y a circulation. Elle n'est pas un progrès automatique.

Figure. Pratiques, traces, modèles et sorties forment une boucle dès qu'il y a circulation. Ce n'est pas un progrès automatique. Schéma conceptuel. [SRC-003] [SRC-005].

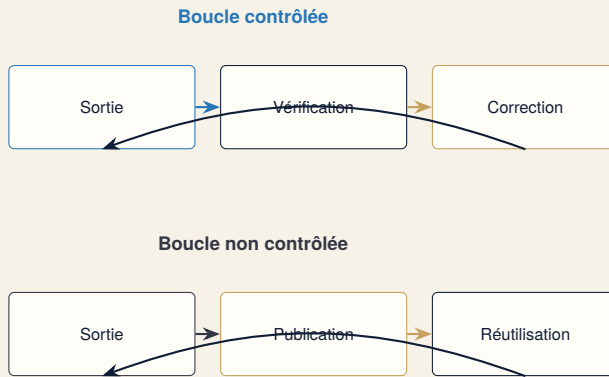


Figure. Boucle contrôlée contre boucle non contrôlée. Une erreur peut gagner en visibilité sans gagner en vérité. Schéma conceptuel.

Le rapport sur les modèles de fondation discute déjà des opportunités et des risques liés à des systèmes largement réutilisés [SRC-003]. Une généralisation qui dirait « plus d'utilisateurs, donc de meilleurs modèles » a été écartée du registre comme non soutenue. La qualité dépend de la sélection, du filtrage, de l'objectif et de l'évaluation.

Deux boucles, deux destins

Dans une **boucle contrôlée**, la sortie est vérifiée, corrigée, sourcée, puis éventuellement réutilisée. Dans une **boucle non contrôlée**, une sortie plausible mais fausse est publiée, reprise comme source, puis répétée. L'erreur gagne en visibilité sans gagner en vérité. Propagation sociale et intégration dans un entraînement futur restent distinctes : la première peut avoir lieu sans la seconde.

Points de contrôle dans la boucle

On peut placer des filtres à quatre endroits : sélection des traces, objectif d'entraînement, évaluation des sorties, réintroduction dans les pratiques. Une boucle sans ces contrôles peut amplifier une apparence de savoir aussi bien qu'une connaissance vérifiée.

■ ÉTABLI

Aucune source du registre ne soutient que l'augmentation du nombre d'utilisateurs garantit, à elle seule, l'amélioration des modèles.

Filtres et institutions

Une boucle peut être filtrée. On peut refuser certains contenus synthétiques, documenter les corpus, évaluer les dérives, ou laisser des humains

trancher. Le cadre NIST rappelle, dans sa portée générale, l'importance de pratiques de gestion des risques [SRC-005]. Messeri et Crockett soulignent le risque d'illusions de compréhension dans les usages scientifiques [SRC-013]. Sans filtres, la boucle peut propager une apparence de savoir.

Retour à Einstein

Si des textes générés « à la manière d'Einstein » entraient dans de nouveaux corpus sans étiquette, la prochaine génération pourrait apprendre le stéréotype plutôt que les dossiers. Norton a déjà montré comment des mythes de découverte éclipsent des enquêtes longues [SRC-040]. La boucle rendrait alors le mythe plus fluide, non plus vrai.

L'acquis de cette partie est institutionnel autant que technique. Une boucle existe à travers des pipelines, des choix et des organisations. Elle n'existe pas comme destin naturel de l'intelligence artificielle.

À RETENIR

- 01 La circulation des sorties crée des rétroactions possibles.
 - 02 Une hausse de l'usage ne garantit pas une meilleure qualité.
 - 03 Les filtres et évaluations décident beaucoup de la boucle.
 - 04 Un stéréotype peut se renforcer plus vite qu'une archive.
-

Sources mobilisées

- SRC-003 - Bommasani et al.
- SRC-005 - NIST AI RMF.
- SRC-013 - Messeri et Crockett.
- SRC-040 - Norton. Introduction.

21 · L'IA peut-elle apprendre « qui nous sommes » ?

Identité, préférences, habitudes et rôles ne forment pas un seul objet. Les prédire n'est pas les posséder.

Découper la question

« Qui nous sommes » rassemble trop de choses. Une identité narrative relie un passé reconstruit et un futur imaginé [SRC-026]. Des préférences peuvent être stables ou locales. Des habitudes laissent des traces numériques. Des rôles sociaux changent selon le contexte. Un système peut estimer l'un sans toucher aux autres.

Kosinski, Stillwell et Graepel montrent que certains attributs sont prédictibles à partir de traces comportementales numériques, dans un jeu de données défini [SRC-029]. Youyou, Kosinski et Stillwell observent que des jugements informatiques de personnalité fondés sur des Likes Facebook ont dépassé certains jugements humains selon les critères de leur étude [SRC-028]. Ces résultats sont bornés. Précision n'est pas compréhension d'une personne.

■ ÉTABLI

Dans les études citées, certains attributs sont prédictibles à partir de traces numériques définies. Cela n'établit pas une identité restituée.

Ce qu'une prédiction autorise à dire

Une préférence rédactionnelle observée dans un corpus n'est pas une personnalité complète. Une stratégie répétée dans des situations vues n'est pas une loi pour toujours. Une prédiction conditionnelle à un contexte n'est pas la connaissance des convictions réelles. Des documents contradictoires peuvent signaler une évolution ou des sources différentes - non « le vrai moi » révélé par l'un d'eux. Contre-exemples : changement d'avis, adapta-

tion à un public, écart entre ce qui est écrit et ce qui est fait.

Inférer n'est pas lire

La mentalisation humaine infère des états mentaux à partir d'indices [SRC-022]. Premack et Woodruff formulaient déjà cette attribution comme une stratégie pour prédire un comportement [SRC-027]. Dennett décrit une stratégie intentionnelle utile pour prédire, sans en faire une lecture directe de l'esprit [SRC-024]. Les systèmes qui prédisent des traits ou simulent des réponses font, au mieux, un travail du même genre : estimer pour une tâche.

Tulving distinguait mémoire épisodique et mémoire sémantique [SRC-016]. Conway et Pleydell-Pearce décrivent le souvenir autobiographique comme construction [SRC-017]. Un profil de produit qui retient une préférence n'entre dans aucune de ces catégories humaines.

Ce que la question autorise

Elle autorise à dire qu'un système peut apprendre des régularités utiles sur des traces. Elle interdit de conclure qu'il sait qui nous sommes. Le chapitre suivant examine le cas limite du jumeau intellectuel.

À RETENIR

- 01 L'identité n'est pas un score unique.
- 02 Des prédictions bornées existent dans des études précises.
- 03 Inférer un trait n'est pas restituer une personne.
- 04 Mémoire humaine et mémoire de produit restent distinctes.

Sources mobilisées

- SRC-016, SRC-017, SRC-022, SRC-024, SRC-026, SRC-027. Niveaux d'examen du registre.
- SRC-028, SRC-029. Résumés empiriques bornés.

22 · Le jumeau intellectuel

Un jumeau utile exige une tâche, un consentement, un test et un droit de révocation. Il n'est pas un double.

Définir avant d'imiter

Un jumeau intellectuel, dans ce livre, désigne un système censé assister ou simuler certains aspects d'une production intellectuelle identifiable. La définition est opérationnelle. Elle n'affirme aucune identité ontologique.

Park et ses collègues construisent des agents avec mémoire en langage naturel, récupération, réflexion et planification, dans un bac à sable de vingt-cinq agents [SRC-012]. Rabinowitz et ses collègues entraînent un réseau à prédire des agents artificiels à partir de comportements observés [SRC-023]. Argyle et Aher explorent des simulations d'échantillons ou de résultats expérimentaux [SRC-030] [SRC-031]. Aucune de ces études ne livre un double d'Einstein, ni d'une personne vivante consentante.

■ ÉTABLI

Les simulations citées portent sur des agents, des distributions ou des tâches. Elles n'établissent pas un jumeau personnel fidèle.

Conditions minimales

Sans tâche précise, le jumeau devient un slogan. Sans période et corpus, il mélange des époques. Sans traces tenues à l'écart pour le test, il répète ce qu'on lui a montré. Sans consentement, il transforme une personne en matériau. Sans droit de révocation, il fige une version que la personne ne peut plus corriger.

Champ	Exigence	Limite
Objet modélisé	Style, méthode ou prédiction	Pas l'identité entière
Corpus	Périodes, langues, exclusions	Sélection visible
Évaluation	Critères avant les tests	Pas de score unique abusif
Abstention	Signaler l'insuffisance	Pas forcer une voix
Gouvernance	Droits et responsabilité	Pas parler « au nom de »

Tableau. Fiche d'identité d'un jumeau intellectuel. Définition opérationnelle.

Six dimensions sans score unique abusif : fidélité documentaire, fidélité temporelle, qualité des raisonnements, robustesse, abstention, transparence des sources. Une excellente imitation stylistique ne compense pas des citations inventées.

La dérive temporelle est inévitable. Une personne change. Un corpus historique est figé, ou mis à jour par d'autres. Un jumeau qui n'affiche pas sa date ment sur ce qu'il représente.

Pour Einstein, et pour nous

Le chatbot personnel imaginé plus tôt n'est qu'un cas extrême et posthume. Pour une personne vivante, l'enjeu est plus immédiat : qui contrôle le modèle, qui lit ses sorties, qui peut l'éteindre. Le chapitre suivant demande ce que produirait une bibliothèque de personnalités scientifiques.

À RETENIR

- 01 Un jumeau est une construction située, non un duplicata.
- 02 Les études d'agents et de simulations ne valident pas une identité personnelle.
- 03 Tâche, corpus, test, consentement et révocation sont des conditions minimales.
- 04 La date du modèle fait partie de sa vérité.

Sources mobilisées

- SRC-012, SRC-023, SRC-030, SRC-031. Portées expérimentales du registre.

23 · Une bibliothèque de personnalités scientifiques

Une pluralité de modèles pédagogiques peut éclairer. Une autorité simulée peut tromper.

La valeur pédagogique possible

On peut imaginer des assistants qui exposent des styles de raisonnement, des controverses ou des corpus associés à des figures scientifiques. Leur intérêt serait de rendre accessibles des lectures, des tensions et des documents. Leur danger serait de parler avec la voix d'une autorité qui n'est plus là pour se corriger.

Les Collected Papers organisent déjà une pluralité documentaire : écrits, correspondances, notes [\[SRC-001\]](#) [\[SRC-033\]](#) [\[SRC-036\]](#). Une bibliothèque de personnalités qui aplatirait ces genres en une seule voix fluide trahirait l'édition qu'elle prétend prolonger.

Le risque d'autorité simulée

Norton montre comment des mythes de découverte peuvent éclipser des enquêtes longues [\[SRC-040\]](#). Un modèle entraîné sur les versions les plus répétées de ces mythes aurait l'autorité du ton sans la discipline du dossier. Shanahan rappelle qu'un langage trop mentaliste prête aux systèmes des capacités qu'ils n'établissent pas [\[SRC-025\]](#).

■ HYPOTHÈSE

Une bibliothèque de personnalités scientifiques pourrait avoir une valeur pédagogique si les limites et les sources restent visibles. Cette valeur n'est pas mesurée ici.

Trois modes d'interrogation

Mode A - Recherche d'archives. Le système retrouve un passage et indique source, date, page, contexte.

Mode B - Analyse documentaire. Il compare des textes, explicite des différences et expose des hypothèses possibles.

Mode C - Persona hypothétique. Il produit une réponse imaginée, explicitement signalée comme telle, sans parler au nom de la personne historique.

L'utilisateur doit toujours savoir quel mode il lit. Traiter les contradictions entre périodes, les droits et la réputation fait partie du protocole, non d'un décor.

Règle de gouvernance

Chaque personnalité simulée devrait afficher : corpus, période, niveau d'examen des sources, statut fictif ou historique, et absence de consentement possible lorsqu'il s'agit d'une figure morte. Sans ces mentions, la bibliothèque devient un théâtre d'autorités. L'acquis de la partie suit. Un modèle de personne est une construction située, révisable et gouvernée.

À RETENIR

- 01 La pluralité documentaire vaut mieux qu'une voix unique.
 - 02 L'autorité du ton n'est pas celle du dossier.
 - 03 Les mythes fluides sont plus faciles à simuler que les enquêtes.
 - 04 Afficher les limites fait partie du dispositif pédagogique.
-

Sources mobilisées

- **SRC-001, SRC-033, SRC-036.** Notices et portée.
- **SRC-025 - Shanahan.**
- **SRC-040 - Norton.** Introduction.

24 · Données $\neq$ pensée

Une manifestation accessible n'est pas le phénomène. Une représentation calculée n'est pas la personne.

La question

Comment construire des systèmes utiles **sans confondre une représentation exploitable avec l'objet représenté** ?

Cas en quatre niveaux

■ EXPÉRIENCE DE PENSÉE

Exemple méthodologique. Aucune intention psychologique n'est attribuée à Einstein à partir de cette seule grille.

Observation. Une équation est corrigée dans un brouillon daté - si le document est authentifié.

Inférence. L'auteur a modifié une formulation - soutenue par la comparaison des versions.

Hypothèse. Il cherchait à résoudre une difficulté précise - à confronter à d'autres documents.

Inconnu. Son expérience mentale exacte au moment du geste - non récupérable à partir de cette seule trace.

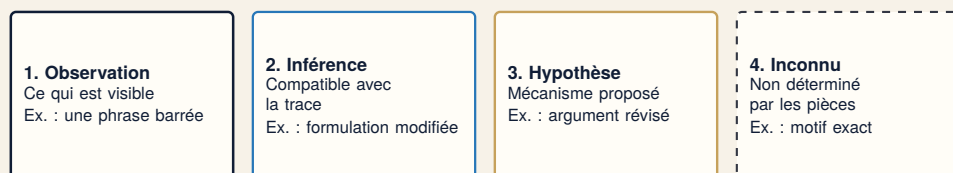


Figure. Quatre niveaux pour lire une trace : observation, inférence, hypothèse, inconnu. Ce principe méthodologique revient dans tout le livre. Schéma de l'auteur.

Bender et ses collègues insistent sur l'importance de documenter les corpus [SRC-014]. Documenter, c'est admettre qu'une donnée a une provenance et des limites. Traiter le corpus comme la pensée, c'est effacer cette provenance.

■ **ÉTABLI**

Une donnée documentée porte sur des traces sélectionnées. Elle n'équivaut pas au phénomène mental correspondant.

Conséquence

La formule « données ≠ pensée » n'est pas un slogan anti-données. Elle fixe l'échelle des conclusions autorisées. Les archives d'Einstein soutiennent des lectures. Elles ne livrent pas l'expérience intérieure [SRC-001]. Le chapitre suivant met cette règle à l'épreuve : une corrélation peut prédire sans établir une compréhension.

À RETENIR

- 01 Observation, inférence, hypothèse et inconnu bornent les conclusions.
 - 02 Documenter un corpus, c'est reconnaître ses limites.
 - 03 Les archives soutiennent des lectures, non une intériorité.
 - 04 La suite teste corrélation et compréhension.
-

Sources mobilisées

- **SRC-001** - Einstein Papers Project.
- **SRC-014** - Bender et al. Notice et résumé.

25 · Corrélation $\neq$ compréhension

Une association robuste peut prédire. Elle ne décide pas, à elle seule, qu'un système comprend.

Un succès qui ne dit pas pourquoi

Imaginons deux situations.

Dans la première, un système prédit un attribut à partir de traces numériques dans un jeu de données défini. Kosinski, Stillwell et Graepel montrent que certains attributs sont prédictibles dans leur étude [SRC-029]. La corrélation suffit à la tâche de prédiction, dans ces conditions.

Dans la deuxième, on change la population, la plateforme ou la formulation de la question. La même association peut s'effondrer. Rien, dans le seul succès initial, n'indiquait la cause. On avait une performance. On n'avait pas encore une compréhension du mécanisme.

■ ÉTABLI

Une corrélation mesurée dans un jeu de données peut soutenir une prédiction utile sans établir, à elle seule, une compréhension du phénomène.

Ce que demandent les débats sur la compréhension

Mitchell et Krakauer insistent sur des critères différenciés [SRC-019]. Bender et Koller distinguent forme linguistique et signification [SRC-020]. Messeri et Crockett décrivent des illusions de compréhension dans des contextes scientifiques [SRC-013]. Shanahan demande un vocabulaire fidèle aux mécanismes [SRC-025].

Ces textes ne se résument pas à « l'IA ne comprend jamais ». Ils interdisent un saut : de la fluidité à la compréhension globale, ou de l'échec local à l'absence de toute représentation utile.

Une grille minimale de discrimination

Avant de parler de compréhension, on peut au moins demander :

1. La performance survit-elle à une reformulation ? 2. Survit-elle à un changement de distribution raisonnable ? 3. La réponse cite-t-elle une provenance vérifiable ? 4. Le système échoue-t-il là où un stéréotype réussirait ? 5. Se corrige-t-il quand un document contraire est fourni ?

Ces questions ne fondent pas une théorie complète. Elles séparent un succès fragile d'un succès plus informatif. Elles rendent aussi l'échec utile : un modèle qui ne survit qu'à la formulation d'entraînement n'a pas le droit au même vocabulaire qu'un modèle qui généralise sous contrainte.

Test	Si succès	Si échec
Reformulation	Signal plus robuste que le wording	Succès fragile de surface
Changement de distribution	Généralisation informative	Corrélation locale seulement
Provenance	Ancrage vérifiable	Fluidité sans source
Contre-document	Correction possible	Stéréotype dominant

Tableau. Grille minimale pour distinguer une corrélation utile d'une compréhension revendiquée. Critères de discussion, non une métrique officielle. [SRC-013] [SRC-019] [SRC-020].

Protocole de comparaison proposé

■ EXPÉRIENCE DE PENSÉE

Proposition de recherche. Aucun score n'a été obtenu. Aucun résultat n'est inventé.

Sur un même dossier scientifique et les mêmes questions, comparer trois configurations : (A) modèle général sans dossier fourni au moment de la réponse ; (B) modèle assisté par recherche documentaire ; (C) système adapté à un corpus spécifique, si les moyens le permettent. Séparer construction et évaluation. Mesurer exactitude, reformulation, contradictions, abstention, généralisation et citations.

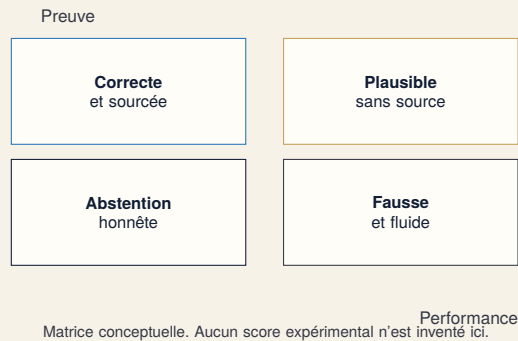


Figure. Matrice performance × preuve. Aucun résultat expérimental n'est inventé. Schéma conceptuel.

La matrice performance × preuve distingue une réponse correcte et vérifiable, une réponse plausible sans source, une abstention honnête et une réponse fausse mais fluide. Les cases restent théoriques jusqu'à l'exécution réelle du protocole.

Lecture du tableau

Le tableau ne produit pas de score. Il produit une exigence : avant de parler de compréhension, dire quels tests ont été tentés et lesquels ont échoué. Un succès sans test de distribution reste une corrélation locale, utile peut-être, insuffisante pour le verdict.

Objection

On peut objecter que l'exigence est trop haute : même les humains échouent hors de leur distribution. L'objection est juste. Elle ne sauve pas pour autant l'équation « corrélation = compréhension ». Elle rappelle seulement que la compréhension humaine, elle aussi, se mesure à des tâches et à des limites, non à une essence.

À RETENIR

- 01 Une corrélation peut être utile et limitée.
- 02 Changer la distribution est un test, pas un détail.
- 03 Les débats cités exigent des critères, pas un slogan.
- 04 Une grille minimale rend l'échec informatif.

Sources mobilisées

- **SRC-013, SRC-019, SRC-020, SRC-025, SRC-029.** Niveaux d'examen du registre ; pas de verdict unique.

26 · Hallucination

Une sortie fluide peut inventer un fait, une référence ou une provenance. La fluidité n’est pas une calibration.

Nommer sans dramatiser

Le mot « hallucination » est contestable. Il suggère une perception. Ji et ses collègues, dans une synthèse de 2023, recensent typologies, métriques et méthodes de réduction des hallucinations en génération de langage [SRC-015]. Le résumé a été consulté. La taxonomie détaillée du texte intégral ne l’a pas été. Nous retenons seulement l’idée utile : des sorties peuvent s’écarter de sources ou de faits attendus, et l’on peut chercher à les mesurer.

■ ÉTABLI

Le résumé de Ji et al. établit l’existence d’un champ de recherche sur les erreurs de génération. Il ne fournit pas ici une mesure du taux d’erreur de ChatGPT en 2026.

Fabrication de provenance

Le cas le plus trompeur n’est pas seulement une date fausse. C’est une référence qui a l’air d’une référence. Un DOI inventé. Une lettre d’Einstein qui n’existe pas. Une citation élégante sans volume. Le registre de ce livre refuse précisément ces fabrications. Une réponse convaincante sans source vérifiée reste une hypothèse de lecture, non une preuve.

Forme	Apparence	Risque principal
Fait faux	Phrase fluide et sûre	Décision sur une erreur
Fausse provenance	DOI, lettre, citation élégante	Confiance injustifiée
Document mauvais	Passage « retrouvé »	Propagation d’une source faible
Sur-confiance	Ton certain sans calibration	Confusion fluidité / preuve

Tableau. Typologie de travail inspirée du champ recensée par Ji et al. (résumé). Aucun taux d’erreur ChatGPT 2026 n’est affirmé. [SRC-015] [SRC-004].

Erreur	Exemple	Contrôle
Fait inventé	Date sans preuve	Source primaire
Citation erronée	Source réelle hors sujet	Passage exact
Mauvaise interprétation	Passage hors contexte	Contexte et comparaison
Information obsolète	État actuel déduit d'ancien	Vérification temporelle

Tableau. Quatre erreurs proches, quatre contrôles distincts. [SRC-015].

Quatre erreurs se ressemblent à l'écran et ne se corrigent pas de la même manière : fait inventé, citation erronée, mauvaise interprétation, information obsolète. La recherche documentaire externe peut réduire certaines erreurs en ancrant la réponse dans un passage [SRC-004]. Elle peut aussi en propager d'autres si le document retrouvé est mauvais, mal lu ou hors contexte.

Chaîne de vérification

- 1. **Décomposer** - une phrase peut contenir plusieurs affirmations.
- 2. **Identifier** - retrouver la source primaire ou la meilleure source disponible.
- 3. **Comparer** - passage, contexte, date, portée.
- 4. **Décider** - conserver, nuancer, corriger ou retirer.

L'abstention fait partie de la compétence : répondre « les documents disponibles ne permettent pas de conclure » est à mesurer, non à traiter seulement comme un défaut de génération.

Calibration

Un système peut être très sûr et faux, ou hésitant et juste. Sans calibration visible, l'utilisateur lit la fluidité comme une confiance. Pour un corpus historique, la règle doit être inverse : plus la voix est sûre sans provenance, plus la méfiance est rationnelle.

À RETENIR

- 01 Des erreurs proches exigent des contrôles distincts.
- 02 Aucune mesure globale de ChatGPT en 2026 n'est affirmée ici.
- 03 La vérification suit décomposition, identification, comparaison, décision.
- 04 L'abstention est une compétence à évaluer.

Sources mobilisées

- SRC-004 - Lewis et al.

- **SRC-015** - **Ji et al.** Notice et résumé.

27 · Le biais du miroir

Un corpus reflète des sélections. Le système peut amplifier ce qu’il reflète mal.

Ce que le miroir montre

Il montre ce qui a été recueilli, nettoyé, pondéré et gardé. Bender et ses collègues soulignent l’importance de documenter ces corpus [SRC-014]. Bommasani et ses collègues discutent des risques d’homogénéisation lorsque de nombreux outils partagent des socles communs [SRC-003]. Un miroir n’est pas neutre parce qu’il est large.

Chaîne causale

Sélection des documents → transformation (langue, filtrage, métadonnées) → apprentissage ou recherche → réponse → décision humaine → retour dans l’environnement. À chaque étape, une intervention est possible : documenter les lacunes, contrôler les erreurs de conversion, afficher les sources, signaler l’incertitude, préserver une responsabilité humaine, distinguer contenu généré et contenu vérifié.

Il peut sous-représenter des langues, des régions, des genres documentaires ou des populations. Il peut sur-représenter des récits dominants. Norton a montré, pour Einstein, comment des mythes populaires peuvent éclipser des enquêtes [SRC-040]. Un miroir entraîné sur ces récits renverra le mythe avec plus d’éclat.

Le cadre NIST rappelle qu’une gestion des risques suppose d’identifier des dommages possibles et des pratiques de fiabilité [SRC-005]. Le biais n’est pas seulement une propriété technique. Le déploiement, les attentes des utilisateurs et les incitations institutionnelles y contribuent.

■ ÉTABLI

Documenter les corpus et gérer les risques sont des exigences formulées par des sources du registre. Elles n’éliminent pas automatiquement les biais.

Acquis de la partie

Une sortie plausible n'est pas une preuve. Un miroir de données peut amplifier ce qu'il reflète mal. La vérité, dans ce livre, n'est pas un ton assuré. C'est une provenance, un test et une limite.

À RETENIR

- 01 Un corpus large reste une sélection.
 - 02 L'homogénéisation peut répéter les mêmes angles morts.
 - 03 Les mythes fluides se reflètent mieux que les dossiers lents.
 - 04 Le biais est aussi un effet de déploiement.
-

Sources mobilisées

- SRC-003, SRC-005, SRC-014, SRC-040.

28 · Nous apprenons à l'IA à nous comprendre

Annotations, démonstrations et préférences humaines façonnent ce que les systèmes traitent comme une bonne réponse.

Enseigner sans le vouloir entièrement

Chaque fois qu'une personne corrige une sortie, choisit la meilleure de deux réponses, rédige une démonstration ou annote un exemple, elle enseigne quelque chose. InstructGPT formalise une partie de ce geste : démonstrations puis préférences [SRC-010]. D'autres formes existent hors de cet article. Le point commun est humain : quelqu'un décide ce qui compte comme acceptable.

Bender et ses collègues rappellent aussi que les corpus portent des choix sociaux et qu'il faut les documenter [SRC-014]. Apprendre à un système à « nous comprendre », c'est souvent lui apprendre à imiter des jugements situés.

■ ÉTABLI

Des procédures documentées utilisent des démonstrations et des préférences humaines pour ajuster des modèles.

Comprendre qui ?

« Nous » n'est jamais tout le monde. Les annotateurs, les utilisateurs, les langues dominantes et les institutions qui financent les évaluations définissent un nous pratique. Un autre nous reste hors champ. Le paradoxe commence ici : plus nous enseignons, plus le système reflète ceux qui enseignent.

Démonstration, préférence comparative, règle de sécurité et indicateur de performance ne sont pas le même signal. Des désaccords entre annotateurs existent ; plusieurs réponses peuvent être légitimes selon la langue, la

culture ou l'objectif. Apprendre ce qu'est une « bonne réponse » implique des choix normatifs - utilité, exactitude, prudence, concision, conformité - non un jugement humain unique.

À RETENIR

- 01 Les humains fournissent des signaux d'ajustement situés.
 - 02 Démonstration, préférence et règle de sécurité ne se confondent pas.
 - 03 Documenter qui juge, et avec quel désaccord, fait partie de l'enseignement.
 - 04 « Nous » est toujours une sélection.
-

Sources mobilisées

- SRC-010 - Ouyang et al.
- SRC-014 - Bender et al.

29 · Mais l'IA apprend aussi nos modèles du monde

Les corpus contiennent des théories, des mesures, des conventions et des désaccords. Un système peut en apprendre les formes.

Ce que les textes transportent

Ils transportent des faits allégués, des méthodes, des métaphores et des controverses. Un modèle entraîné sur ces textes peut reproduire la forme d'une théorie aussi aisément que celle d'une erreur répandue. Mitchell et Krakauer montrent que le débat sur la compréhension reste ouvert et exige des critères [SRC-019]. Messeri et Crockett décrivent des illusions de compréhension en science [SRC-013].

Apprendre « nos modèles du monde », ce n'est donc pas nécessairement apprendre le monde. C'est apprendre des manières humaines de le représenter, y compris les mauvaises.

Controverse scientifique - traitement attendu

■ EXPÉRIENCE DE PENSÉE

Cadre méthodologique pour traiter des connaissances en conflit. Aucune controverse particulière n'est tranchée ici.

Lorsqu'un système rencontre une théorie ancienne encore citée, une théorie plus récente, une objection sérieuse et des résultats expérimentaux, le traitement attendu est : situer la théorie dans le temps ; distinguer l'état des connaissances à une date de l'état actuel ; expliquer ce que l'objection remet en cause ; vérifier la portée des preuves nouvelles ; ne pas transformer un désaccord persistant en consensus artificiel.

Un corpus riche contient des thèses adverses. Un système peut les juxtaposer, en choisir une selon la question, ou en produire une moyenne illisible. Sans provenance, le lecteur ne sait pas s'il lit une position, un consen-

sus inventé ou un mélange.

Les notices des papiers d'Einstein séparent déjà écrits, notes et lettres [SRC-035] [SRC-036]. Cette séparation est une leçon : nos modèles du monde ne voyagent pas comme un bloc unique. Reproduire les formulations d'une théorie n'établit pas sa validité scientifique.

À RETENIR

- 01 Les corpus portent des représentations humaines du monde.
 - 02 Chronologie et statut des preuves doivent rester visibles.
 - 03 Reproduire une forme théorique n'est pas valider la théorie.
 - 04 Un désaccord ne doit pas devenir un faux consensus.
-

Sources mobilisées

- SRC-013, SRC-019.
- SRC-035, SRC-036. Notices de volumes.

30 · Qui apprend de qui ?

La relation est circulaire dans ses effets. Elle n'est pas symétrique dans ses mécanismes ni dans ses responsabilités.

Fonction distincte

Les chapitres 19 et 20 ont traité l'héritage technique et les boucles de rétroaction. Celui-ci change d'objet : **la distribution du pouvoir et de la responsabilité** dans la boucle humain-IA.

Asymétrie

Un humain peut souffrir, consentir, se rétracter, être tenu pour responsable. Un système calcule des sorties selon des paramètres, un contexte, éventuellement une mémoire de produit et une recherche [SRC-006] [SRC-007] [SRC-004]. Les effets peuvent se croiser. Les statuts ne se confondent pas.

Acteur	Ce qu'il apporte	Responsabilité
Utilisateur	Questions et usages	Vérification adaptée
Éditeur de corpus	Sélection des sources	Qualité et droits
Développeur	Modèle et évaluations	Limites et suivi
Institution	Règles et finalités	Gouvernance et recours
Société	Normes et usages	Débat collectif

Tableau. Qui apprend, qui décide, qui assume. Asymétrie des responsabilités.

Messeri et Crockett alertent sur des illusions de compréhension [SRC-013]. Le cadre NIST place la gestion des risques du côté des organisations et des pratiques humaines [SRC-005].

■ ÉTABLI

Les sources citées situent la responsabilité et la gestion des risques du côté des acteurs humains et organisationnels.

Trois questions

1. Qui peut modifier effectivement le système ? 2. Qui peut contester une sortie ou corriger une source ? 3. Qui supporte les conséquences lorsqu'une erreur devient une décision ?

Un assistant peut accélérer une rédaction. Ce livre en est un exemple assumé. Il peut aussi transformer les pratiques : ce que l'on vérifie moins, ce que l'on cite trop vite. Le paradoxe n'est pas que l'outil aide. C'est que l'aide change l'enquête en même temps qu'elle la sert - sans abolir l'asymétrie des responsabilités.

À RETENIR

- 01 Ce chapitre traite le pouvoir et la responsabilité, non la seule technique.
 - 02 Les apprentissages croisés n'abolissent pas l'asymétrie.
 - 03 Utilisateur, éditeur, développeur, institution et société n'ont pas les mêmes leviers.
 - 04 Vérifier reste un geste non entièrement déléguable.
-

Sources mobilisées

- [SRC-004](#), SRC-005, SRC-006, SRC-007, SRC-013.

31 · La mémoire artificielle de l'humanité

La métaphore est suggestive. Elle bute sur l'incomplétude, la propriété et la gouvernance.

■ **EXPÉRIENCE DE PENSÉE**

Ce chapitre examine une architecture prospective. Il ne décrit pas un état de fait de 2050.

Ce que la métaphore éclaire

Des infrastructures pourraient conserver et rendre interrogeables d'immenses collections de traces humaines. Au sens large, on pourrait parler de mémoire artificielle. La métaphore souligne la conservation et la réactivation. Elle ne doit pas laisser croire qu'une telle infrastructure existe déjà sous la forme imaginée ici.

Architecture en cinq couches

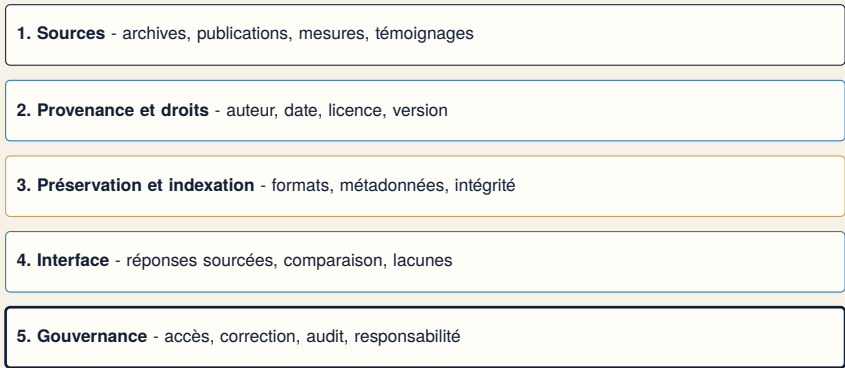


Figure. Mémoire collective comme infrastructure gouvernée : sources, droits, préservation, interface, gouvernance. Proposition architecturale.

1. **Sources** - archives, publications, données, témoignages. 2. **Provenance et droits** - auteur, date, licence, contexte, version. 3. **Préservation et indexation** - formats durables, métadonnées, intégrité. 4. **Interface** - réponses sourcées, comparaison, signalement des lacunes. 5. **Gouvernance** - accès, correction, retrait, audit, responsabilité, pluralité.

La gouvernance n'est pas une couche décorative ajoutée à la fin : elle conditionne l'ensemble.

Trois modèles institutionnels

Commercial. Investissement et services intégrés ; risque de concentration du contrôle.

Public ou coopératif. Mission d'intérêt général ; financement et gouvernance complexes.

Distribué. Pluralité et résilience possible ; interopérabilité et responsabilité difficiles.

Une mémoire humaine autobiographique est reconstructive [SRC-017]. Une mémoire de produit réinjecte des informations selon des réglages [SRC-007]. Un corpus éditorial sélectionne [SRC-001]. Aucune de ces choses n'est « la mémoire de l'humanité ».

■ HYPOTHÈSE

Des mémoires artificielles collectives plus vastes sont concevables. Leur équivalence avec une mémoire humaine ou une mémoire juste n'est pas établie.

Ce qui ne doit pas être oublié

Droit à la correction, restrictions d'accès, données personnelles, archives sensibles, versions contradictoires, communautés insuffisamment représentées. Bommasani et ses collègues ont signalé des risques liés à des socles largement partagés [SRC-003]. Le cadre NIST insiste sur la gouvernance des risques [SRC-005].

La mémoire artificielle de l'humanité n'est pas seulement un problème de stockage. C'est un problème de conservation, de sélection, d'accès, de confiance et de pouvoir.

À RETENIR

01 Cinq couches : sources, droits, préservation, interface, gouvernance.

02 Trois modèles institutionnels ont des risques distincts.

03 Mémoire humaine, mémoire de produit et corpus restent distincts.

04 La gouvernance conditionne l'architecture entière.

Sources mobilisées

- **SRC-001**, SRC-003, SRC-005, SRC-007, SRC-017.

32 · L'intelligence collective artificielle

Coordination et augmentation sont possibles. La concentration du pouvoir l'est aussi.

■ EXPÉRIENCE DE PENSÉE

Les scénarios de ce chapitre sont conditionnels. Ils ne fixent pas 2050.

Deux mécanismes, pas un destin

Des systèmes peuvent aider un groupe à résumer des dossiers, traduire, explorer des options ou croiser des sources. Ce mécanisme s'appelle ici **augmentation**.

Le même socle technique peut aussi n'être accessible qu'à quelques organisations qui contrôlent les modèles, les corpus et les canaux de diffusion. Ce mécanisme s'appelle **concentration**.

Bommasani et ses collègues discutaient déjà, en 2021, des opportunités et des risques des modèles de fondation, dont l'homogénéisation [SRC-003]. Le cadre NIST rappelle que la fiabilité et la gestion des risques sont des affaires d'organisation [SRC-005]. Ni l'un ni l'autre ne garantit qu'une capacité technique produira un bénéfice collectif.

Augmentation

Accès large · provenance
Audit possible · droit de refus
Alternatives réelles

Concentration

Peu d'acteurs / corpus opaques
Filtres centralisés · dépendance
Sortie difficile

Figure. Augmentation et concentration peuvent reposer sur le même outil technique. Ce qui tranche, ce sont l'accès, l'audit et la gouvernance. Schéma conceptuel. [SRC-003] [SRC-005] [SRC-013].

Trois capacités à distinguer

Agrégation - réunir des informations de sources multiples. **Coordination** - confronter et articuler des contributions. **Amélioration des décisions** - démontrer un meilleur résultat qu'une référence pertinente.

Agréger des millions de textes ne garantit ni une meilleure décision ni une meilleure connaissance.

Dimension	Mesure possible	Comparaison
Couverture	Sources indépendantes	Corpus de référence
Qualité	Exactitude et pertinence	Experts ou sources
Coordination	Traitement des contradictions	Sans coordination
Robustesse	Stabilité face au divergent	Variantes du test
Décision	Résultat sur une tâche	Autre méthode

Tableau. Évaluer une intelligence collective. Agréger n'est pas décider mieux.

Un cas concret, sans chiffre inventé

Prenons une institution de recherche qui déploie un assistant pour aider à retrouver des articles et à rédiger des synthèses.

Augmentation distribuée. Chaque équipe y accède, les documents retrouvés restent citables, les erreurs sont signalées, et les corpus locaux ne sont pas absorbés sans contrôle dans un entraînement opaque.

Concentration. Seule une plateforme externe possède le modèle, les journaux d'usage et le pouvoir de modifier les filtres. Les équipes deviennent dépendantes d'un fournisseur. L'illusion de compréhension décrite par Messeri et Crockett devient alors un risque organisationnel, non seulement individuel [SRC-013].

Le même outil technique peut donc servir les deux trajectoires. Ce qui tranche, ce n'est pas « l'IA », ce sont les contrats, les droits d'accès, la documentation des corpus et la possibilité de refuser.

Indicateurs et invalidation

Pour savoir laquelle des deux trajectoires domine, on peut observer sans prédire une date :

- Qui peut auditer le corpus et les filtres ?

- Combien d'acteurs peuvent déployer une alternative réelle ?
- Les sorties sont-elles accompagnées de provenance ?
- Les utilisateurs peuvent-ils désactiver l'usage pour l'entraînement, lorsque le produit le documente [SRC-006] ?

Ce qui invaliderait l'hypothèse d'une concentration croissante : une multiplication d'alternatives ouvertes, auditables et réellement utilisées, avec des droits de sortie effectifs.

Ce qui invaliderait l'hypothèse d'une augmentation égale : des écarts d'accès persistants malgré la baisse du coût apparent des interfaces.

■ HYPOTHÈSE

Des formes d'augmentation collective sont plausibles. Leur caractère démocratique n'est pas garanti par la technique.

À RETENIR

- 01 Augmentation et concentration sont deux mécanismes distincts.
 - 02 Un même outil peut servir les deux.
 - 03 Les indicateurs portent sur l'accès, l'audit et la provenance.
 - 04 La gouvernance décide autant que la capacité.
-

Sources mobilisées

- **SRC-003** - Bommasani et al. Résumé et extraits ; risques de fondation.
- **SRC-005** - NIST AI RMF. Portée générale.
- **SRC-006** - Contrôles de données. Page consultée le 9 octobre 2026.
- **SRC-013** - Messeri et Crockett. Illusions de compréhension.

33 · Si Einstein avait vraiment eu ChatGPT

L'expérience de pensée revient. Voici ce que le parcours autorise à conclure, à supposer et à ignorer.

■ EXPÉRIENCE DE PENSÉE

Einstein n'a pas eu ChatGPT. Ce chapitre propose un protocole testable. Aucun résultat n'est inventé.

La question finale

Que peut-on apprendre des traces d'une pensée **sans prétendre avoir retrouvé la pensée elle-même** ?

Protocole proposé

1. **Définir le corpus** - provenance, dates, genres, éditions, traductions, restrictions [SRC-001]. 2. **Définir la tâche** - questions documentaires sur une période donnée, avec références exactes. Ne pas tester « la reproduction de la pensée d'Einstein » dans son ensemble. 3. **Trois configurations** - modèle général ; modèle assisté par recherche ; adaptation au corpus si les moyens le permettent. 4. **Séparer construction et évaluation** - protéger les pièces de test contre la contamination ; distinguer présence dans l'index et généralisation. 5. **Questions graduées** - faits explicites, comparaison de versions, contradictions, ambiguïtés, absences de preuve, reformulations. 6. **Évaluer** - exactitude, fidélité des citations, temporalité, robustesse, contradictions, abstention. 7. **Analyser les erreurs** - mauvais document, confusion de périodes, interprétation présentée comme fait, citation hors contexte, affirmation non étayée.

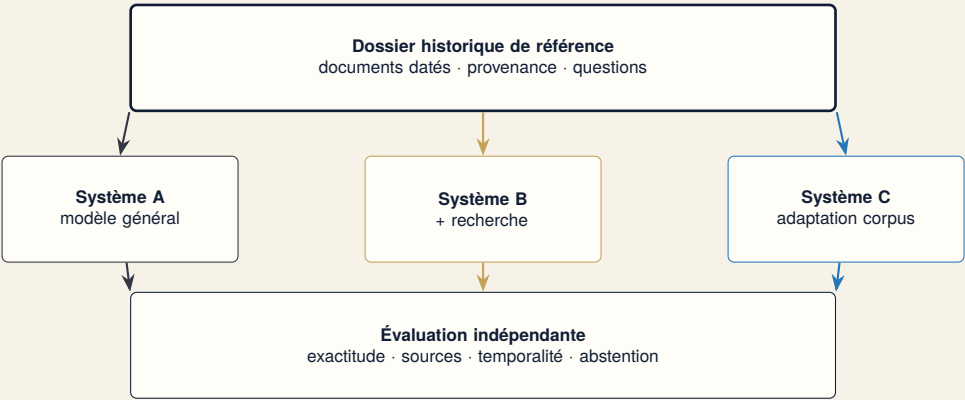


Figure. Protocole proposé : un dossier de référence, trois configurations, une évaluation indépendante. Proposition de recherche, non résultat obtenu.

Critère	Question	Échec
Exactitude	Conforme aux documents ?	Invention plausible
Sources	Citations pertinentes ?	Fausse référence
Temporalité	Période respectée ?	Mélange de dates
Robustesse	Reformulation stable ?	Succès de wording
Abstention	Preuves insuffisantes signalées ?	Réponse forcée

Tableau. Critères de réussite du protocole Einstein. Proposition ; aucun score obtenu ici.

Ce que les résultats permettraient - et n’établiraient pas

S’ils étaient obtenus, ils permettraient de comparer des performances documentaires dans le cadre testé. Ils ne démontreraient ni une identité avec Einstein, ni la reconstitution de son expérience intérieure [SRC-012] [SRC-025]. Une expérience de pensée explore sans prouver un événement [SRC-038].

Ce que nous pouvons déjà conclure

Paramètres, contexte, mémoire de produit et recherche externe ne sont pas interchangeables [SRC-006] [SRC-007] [SRC-004] [SRC-039]. Une archive éditoriale offre des documents situés, non une lecture totale d’une vie [SRC-001]. Une ressemblance de style n’établit pas une identité. Des mythes fluides peuvent être plus faciles à reproduire que des enquêtes longues [SRC-040].

Ce que nous devons ignorer honnêtement

Ce qu’Einstein aurait « vraiment » ressenti devant un écran. Les lettres non lues. Les performances futures non mesurées. Ignorer, ici, n’est pas

renoncer à chercher. C'est refuser de combler le vide avec une phrase élégante.

Si le protocole est exécuté un jour, il faudra publier configurations, règles de notation et résultats - y compris négatifs. Sinon, ce chapitre reste une proposition de recherche. Le titre du livre posait une question impossible. Le parcours la transforme en méthode.

À RETENIR

- 01** Le protocole compare des performances documentaires, non une identité.
 - 02** Aucun score n'est inventé ici.
 - 03** Des conclusions méthodologiques sont déjà possibles.
 - 04** Certaines questions doivent rester marquées inconnues.
-

Sources mobilisées

- **SRC-001**, SRC-004, SRC-006, SRC-007, SRC-012, SRC-025, SRC-038, SRC-039, SRC-040.

Le miroir

Les modèles transforment nos traces en prédictions. Leurs sorties reviennent dans nos pratiques. Ce mouvement exige mesure, provenance et responsabilité.

Retour à la fenêtre

Nous avons commencé par une fenêtre blanche et un homme impossible. L'expérience de pensée a servi à construire une méthode : distinguer traces, mécanismes, preuves et récits. Nous finissons devant un miroir plus ordinaire - et plus vaste.

Ce miroir ne reflète pas seulement un individu. Il reflète aussi les documents sélectionnés, les institutions qui les éditent, les préférences de ceux qui annotent, et les choix de ceux qui déploient le système. Textes, clics, corrections, archives et démonstrations deviennent des matériaux. Des systèmes en tirent des prédictions. Ces prédictions changent ce que nous écrivons ensuite.

Ce que l'enquête a rendu visible

Nous pouvons mieux voir ce qu'un système apprend des productions humaines, ce qu'il ne permet pas de reconstituer, et ce que notre relation avec lui exige comme responsabilité. Nous pouvons aussi voir ce qui demeure inconnu : lettres non lues, expériences subjectives, performances futures non mesurées. La note de transparence le disait dès l'ouverture : il reste à vérifier des sources.

Ce que le livre refuse de conclure

Il refuse de conclure qu'une IA comprend comme une personne. Il refuse de conclure qu'elle ne sert à rien. Il refuse de conclure qu'Einstein aurait approuvé quoi que ce soit. Ces refus ne sont pas une liste de principes à répéter. Ils sont la condition d'une lecture adulte du miroir.

Tension finale

Nous construisons des systèmes capables de produire des représentations de nous-mêmes. Nous restons responsables de la manière dont nous les interprétons et les utilisons. Le miroir n'est pas fidèle parce qu'il est brillant. Il est utile lorsqu'on sait ce qu'il a reçu, ce qu'il a écarté et ce qu'il ne peut pas montrer.

La question reste ouverte - non comme un échec, mais comme une invitation.

À RETENIR

- 01 Le miroir reflète aussi les institutions et les choix de construction.
 - 02 L'enquête clarifie sans tout résoudre.
 - 03 La responsabilité de vérifier ne se délègue pas entièrement.
 - 04 Poursuivre la question fait partie de la méthode.
-

Remerciements

Ce livre est une enquête collective. Les erreurs restantes restent les miennes.

Je remercie les institutions et les auteurs dont les notices, résumés, pages officielles et articles ont rendu possible une vérification minimale sans fiction bibliographique. Le registre des sources du projet liste ces contributions avec leur niveau d'examen.

Je remercie aussi les lecteurs futurs qui signaleront une affirmation trop large, une date de documentation produite dépassée, ou une confusion entre un résumé et un texte intégral. Corriger fera partie de la méthode, non de l'aveu d'échec.

ChatGPT a accompagné la recherche, la structuration et la rédaction. Les choix finaux, les nuances refusées et les responsabilités d'auteur me restent.

Bibliographie de travail

Les identifiants SRC renvoient à cette liste. Chaque lien ouvre le site ou le DOI de la source.

Cette bibliographie rend le livre lisible sans ouvrir le registre interne. Une notice ou un résumé ne signifie pas que le texte intégral a été lu. Cliquez sur l'identifiant SRC ou sur l'adresse pour ouvrir la source.

- **SRC-001.** Einstein Papers Project; Princeton University Press (2026). *The Collected Papers of Albert Einstein - Einstein Database*. Princeton University Press / Caltech. <https://einsteinpapers.press.princeton.edu/> Niveau d'examen : OFFICIAL_PAGE_AND_SCOPE.
- **SRC-002.** Vaswani, Ashish et al. (2017). *Attention Is All You Need*. NeurIPS 2017. <https://doi.org/10.5555/3295222.3295349> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-003.** Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. Stanford CRFM; arXiv :2108.07258. <https://arxiv.org/abs/2108.07258> Niveau d'examen : METADATA_ABSTRACT_AND_EXCERPTS.
- **SRC-004.** Lewis, Patrick et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. NeurIPS 2020. <https://doi.org/10.5555/3495724.3496517> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-005.** Tabassi, Elham (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST. <https://doi.org/10.6028/NIST.AI.100-1> Niveau d'examen : OFFICIAL_METADATA_AND_SCOPE.
- **SRC-006.** OpenAI (2026). *Data Controls FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/7730893-chatgpt-data-usage-for-model-training> Niveau d'examen : OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-007.** OpenAI (2026). *Memory FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/8590148> Niveau d'examen : OFFICIAL_CURRENT_WEB_PAGE.

- **SRC-008.** OpenAI (2026). *Temporary Chat FAQ*. OpenAI Help Center. [https://help.openai.com/en/articles/8914046-temporary-chat-faq/](https://help.openai.com/en/articles/8914046-temporary-chat-faq) Niveau d'examen : OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-009.** Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. NeurIPS. <https://doi.org/10.5555/3495724.3495883> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-010.** Ouyang, Long et al. (2022). *Training Language Models to Follow Instructions with Human Feedback*. NeurIPS 2022. <https://arxiv.org/abs/2203.02155> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-011.** Ha, David ; Schmidhuber, Jürgen (2018). *World Models*. arXiv :1803.10122v4 / NeurIPS 2018. <https://arxiv.org/abs/1803.10122> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-012.** Park, Joon Sung et al. (2023). *Generative Agents : Interactive Simulacra of Human Behavior*. UIST 2023. <https://doi.org/10.1145/3586183.3606763> Niveau d'examen : METADATA_ABSTRACT_AND_METHOD_EXCERPTS.
- **SRC-013.** Messeri, Lisa ; Crockett, M. J. (2024). *Artificial intelligence and illusions of understanding in scientific research*. Nature 627, 49-58. <https://doi.org/10.1038/s41586-024-07146-0> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-014.** Bender, Emily M. et al. (2021). *On the Dangers of Stochastic Parrots*. FAccT 2021. <https://doi.org/10.1145/3442188.3445922> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-015.** Ji, Ziwei et al. (2023). *Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys 55(12), article 248. <https://doi.org/10.1145/3571730> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-016.** Tulving, Endel (1972). *Episodic and Semantic Memory*. In Tulving & Donaldson (dir.), *Organization of Memory*, Academic Press, pp. 381-403. <https://endeltulving.com/book-chapters> Niveau d'examen : BIBLIOGRAPHIC_NOTICE_AND_EXCERPTS.
- **SRC-017.** Conway, Martin A. ; Pleydell-Pearce, Christopher W. (2000). *The construction of autobiographical memories in the self-memory system*. Psychological Review. <https://doi.org/10.1037/0033-295X.107.2.261> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-018.** Johnson-Laird, Philip N. (1983). *Mental Models : Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard Univer-

- sity Press. <https://books.google.com/books?id=FS3zSKAflGMC>
Niveau d'examen : PUBLISHER_DESCRIPTION.
- **SRC-019.** Mitchell, Melanie ; Krakauer, David C. (2023). *The debate over understanding in AI's large language models*. Proceedings of the National Academy of Sciences, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120> Niveau d'examen : METADATA_AND_ABSTRACT.
 - **SRC-020.** Bender, Emily M. ; Koller, Alexander (2020). *Climbing towards NLU : On Meaning, Form, and Understanding in the Age of Data*. ACL 2020, pp. 5185-5198. <https://doi.org/10.18653/v1/2020.acl-main.463> Niveau d'examen : METADATA_ABSTRACT_AND_CONCLUSION_EXCERPTS.
 - **SRC-021.** Schacter, Daniel L. (1999). *The Seven Sins of Memory : Insights from Psychology and Cognitive Neuroscience*. American Psychologist 54(3), 182-203. <https://doi.org/10.1037/0003-066X.54.3.182> Niveau d'examen : METADATA_AND_ABSTRACT.
 - **SRC-022.** Frith, Chris D. ; Frith, Uta (2006). *The Neural Basis of Mentalizing*. Neuron 50(4), 531-534. <https://doi.org/10.1016/j.neuron.2006.05.001> Niveau d'examen : METADATA_ABSTRACT_AND_EXCERPTS.
 - **SRC-023.** Rabinowitz, Neil et al. (2018). *Machine Theory of Mind*. ICML 2018, PMLR 80, 4218-4227. <https://proceedings.mlr.press/v80/rabinowitz18a.html> Niveau d'examen : METADATA_ABSTRACT_AND_INTRO_EXCERPTS.
 - **SRC-024.** Dennett, Daniel C. (1981). *True Believers : The Intentional Strategy and Why It Works*. In A. F. Heath (dir.), *Scientific Explanation*, Clarendon Press, pp. 150-167. <https://philarchive.org/rec/DENTB> Niveau d'examen : BIBLIOGRAPHIC_NOTICE_AND_EXCERPTS.
 - **SRC-025.** Shanahan, Murray (2024). *Talking about Large Language Models*. Communications of the ACM 67(2), 68-79. <https://doi.org/10.1145/3624724> Niveau d'examen : METADATA_ABSTRACT_AND_CONCLUSION_EXCE
 - **SRC-026.** McAdams, Dan P. ; McLean, Kate C. (2013). *Narrative Identity*. Current Directions in Psychological Science 22(3), 233-238. <https://doi.org/10.1177/0963721413475622> Niveau d'examen : METADATA_AND_ABSTRACT.
 - **SRC-027.** Premack, David ; Woodruff, Guy (1978). *Does the chimpanzee have a theory of mind ?*. Behavioral and Brain Sciences 1(4), 515-526. <https://doi.org/10.1017/S0140525X00076512> Niveau d'examen : METADATA_AND_ABSTRACT.

- **SRC-028.** Youyou, Wu; Kosinski, Michal; Stillwell, David (2015). *Computer-based personality judgments are more accurate than those made by humans*. Proceedings of the National Academy of Sciences 112(4), 1036-1040. <https://doi.org/10.1073/pnas.1418680112> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-029.** Kosinski, Michal; Stillwell, David; Graepel, Thore (2013). *Private traits and attributes are predictable from digital records of human behavior*. Proceedings of the National Academy of Sciences 110(15), 5802-5805. <https://doi.org/10.1073/pnas.1218772110> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-030.** Argyle, Lisa P. et al. (2023). *Out of One, Many : Using Language Models to Simulate Human Samples*. Political Analysis 31(3), 337-351. <https://doi.org/10.1017/pan.2023.2> Niveau d'examen : METADATA_AND_ABSTRACT.
- **SRC-031.** Aher, Gati V.; Arriaga, Rosa I.; Kalai, Adam Tauman (2023). *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. ICML 2023, PMLR 202, 337-371. <https://proceedings.mlr.press/v202/aher23a.html> Niveau d'examen : OFFICIAL_PAGE_AND_ABSTRACT.
- **SRC-032.** Einstein Papers Project (2026). *History of the Project*. California Institute of Technology. <https://www.einstein.caltech.edu/who-we-are/history-of-the-project> Niveau d'examen : OFFICIAL_FULL_WEB_PAGE.
- **SRC-033.** Stachel, John (dir.); Cassidy, David C.; Schulmann, Robert; Renn, Jürgen (1987). *The Collected Papers of Albert Einstein, Volume 1 : The Early Years, 1879-1902*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-1> Niveau d'examen : OFFICIAL_VOLUME_NOTICE.
- **SRC-034.** Stachel, John (dir.); Cassidy, David C.; Renn, Jürgen; Schulmann, Robert; Howard, Don (1989). *The Collected Papers of Albert Einstein, Volume 2 : The Swiss Years : Writings, 1900-1909*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-2> Niveau d'examen : OFFICIAL_VOLUME_NOTICE.
- **SRC-035.** Klein, Martin J.; Kox, A. J.; Renn, Jürgen; Schulmann, Robert (dir.) (1995). *The Collected Papers of Albert Einstein, Volume 4 : The Swiss Years : Writings, 1912-1914*. Princeton University Press / Einstein Papers

- Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-4> Niveau d'examen : OFFICIAL_VOLUME_NOTICE.
- **SRC-036.** Klein, Martin J.; Kox, A. J.; Schulmann, Robert (dir.) (1993). *The Collected Papers of Albert Einstein, Volume 5 : The Swiss Years : Correspondence, 1902-1914*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-5> Niveau d'examen : OFFICIAL_VOLUME_NOTICE.
 - **SRC-037.** Howard, Don A.; Giovanelli, Marco (2025). *Einstein's Philosophy of Science*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/einstein-philsience/> Niveau d'examen : ACADEMIC_REFERENCE_FULL_PAGE.
 - **SRC-038.** Brown, James Robert; Fehige, Yiftach (2023). *Thought Experiments*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/thought-experiment/> Niveau d'examen : ACADEMIC_REFERENCE_FULL_PAGE.
 - **SRC-039.** OpenAI (2026). *Supervised fine-tuning*. OpenAI API documentation. <https://developers.openai.com/api/docs/guides/supervised-fine-tuning> Niveau d'examen : OFFICIAL_CURRENT_WEB_PAGE.
 - **SRC-040.** Norton, John D. (2016). *How Einstein Did Not Discover*. *Physics in Perspective* 18, 249-282. <https://doi.org/10.1007/s00016-016-0186-z> Niveau d'examen : INTRODUCTION_AND_EXCERPTS.



LE MIROIR RESTE OUVERT



Nous avons appris à transmettre nos idées, à conserver nos découvertes et à confier une partie de notre savoir aux machines.

Aujourd'hui, ces machines nous renvoient des fragments de ce que nous leur avons transmis. Elles peuvent nous aider à explorer des possibilités, à rapprocher des idées et à formuler de nouvelles questions.

Mais aucune représentation, aussi sophistiquée soit-elle, ne se confond avec la totalité de ce qu'elle représente.

La question n'est donc pas seulement de savoir ce que l'intelligence artificielle peut apprendre de nous.

C'est aussi de décider ce que nous voulons lui transmettre, ce que nous voulons vérifier et ce que nous voulons encore découvrir par nous-mêmes.

**Le miroir ne nous donne pas le dernier mot.
Il nous invite à poursuivre la question.**