



HORS-SÉRIE

Intelligence artificielle et souveraineté

Capital, calcul, modèles ouverts : ce que la France et l'Europe contrôlent réellement — neuf thèses falsifiables et leurs conditions de réfutation

QUESTION DE RECHERCHE

« De quoi la France et l'Europe disposent-elles réellement en matière d'intelligence artificielle, et que perdraient-elles si une dépendance était interrompue ? »

Intelligence artificielle et souveraineté

Capital, calcul, modèles ouverts : ce que la France et l'Europe contrôlent réellement — neuf thèses falsifiables et leurs conditions de réfutation

« De quoi la France et l'Europe disposent-elles réellement en matière d'intelligence artificielle, et que perdraient-elles si une dépendance était interrompue ? »

STRUCTURE DU DOCUMENT

1. Partie I — Synthèse méthodologique

Synthèse méthodologique · Intelligence artificielle et souveraineté

2. Partie II — Rapport de fond et thèses Nexus

Rapport de fond · Intelligence artificielle et souveraineté

3. Partie III — Recherche empirique source

Rapport de recherche empirique · Intelligence artificielle et souveraineté

STATUT ÉDITORIAL

Publication Nexus officielle. Hors-série méta-rapport transversal : la souveraineté en intelligence artificielle n'entre pas dans le découpage habituel des quatre domaines Nexus, elle les traverse tous — économie, société, infrastructures, éducation. Le document réunit trois livrables complémentaires — la synthèse méthodologique (cadrage, cartographie du débat, état empirique, controverses, angles morts, cahier des sources), le rapport de fond avec les neuf thèses Nexus (THESE-HS-2026-010 à 018) et leurs conditions de réfutation, et la recherche empirique source qui en constitue la matière première. Conformément à la Charte Nexus, chaque thèse reçoit un identifiant unique, un statut épistémique explicite et des conditions de réfutation observables ; quatre thèses sont consolidées, cinq restent des hypothèses de travail.

Référence HORS-SERIE_IA_SOVERAINETE — v2.0

Version 2-0 — validée par le comité de lecture Nexus

Éditeur Nexus, think tank falsificationniste — association loi 1901

Protocole Charte Nexus — sources vérifiées et hiérarchisées, comité de lecture, statut épistémique explicite, révision déclenchée par les conditions de réfutation (mesures de la performance des grappes de calcul, répartition du capital et des droits de vote, mesures de distribution des modèles ouverts, répartition des marchés publics par niveau d'assurance, taux d'adoption par taille d'entreprise et de collectivité).

Sommaire

PARTIE I — SYNTHÈSE MÉTHODOLOGIQUE

I. Cadrage : la question posée	6
II. Cartographie du débat	7
III. État empirique	9
IV. Controverses	13
V. Angles morts	15
VI. Les neuf thèses falsifiables — rappel structurel	16
VII. Cahier des sources	17

PARTIE II — RAPPORT DE FOND ET THÈSES NEXUS

Résumé exécutif	20
Préambule	22
I. Cadrage : ce que « souveraineté » veut dire en intelligence artificielle	23
II. État empirique : la géographie du calcul	25
III. Le capital : la levée de fonds Mistral AI de septembre 2026	28
IV. Les modèles ouverts : un basculement vers la Chine	31
V. Le repli de l'ouverture	33
VI. Cartographie des acteurs et du débat	35
VII. Une distinction structurante : conformité et capacité	36
VIII. Le cadre européen	38
IX. Les positions institutionnelles françaises	40
X. Les autres puissances	41
XI. Les neuf thèses falsifiables	42
XII. Implications pour les politiques publiques	49
XIII. Implications pour les territoires	50
XIV. Angles morts	51
XV. Trois scénarios à l'horizon 2032	52
XVI. Bibliographie hiérarchisée	53
XVII. Glossaire	58

PARTIE III — RECHERCHE EMPIRIQUE SOURCE

I. Périmètre et enjeux	61
II. Méthodologie et hiérarchie des sources	61
III. La levée de fonds Mistral AI de septembre 2026	62
IV. La bascule de l'open weights vers la Chine	65
V. Le repli de l'ouverture : acteurs américains et signaux chinois	67
VI. La géopolitique du calcul	69
VII. Le cadre européen	71
VIII. Les positions institutionnelles françaises	74
IX. Les sommets internationaux	77
X. La demande : adoption en France et dans les territoires	78
XI. Tableau comparatif des dépendances par couche	80
XII. Sources et bibliographie sélective	80

PARTIE I

Synthèse méthodologique

SYNTHÈSE MÉTHODOLOGIQUE

Intelligence artificielle et souveraineté

*Cadrage, cartographie du débat, état empirique, controverses, angles morts
— hors-série transversal*

Domaine : Hors-série

Format : Synthèse méthodologique (Charte VI.4)

Version 1-0 — Validée par le comité de lecture Nexus

I. Cadrage : la question posée

I.1 Pourquoi ce hors-série

L'intelligence artificielle est passée, en moins de quatre ans, du statut d'objet de recherche à celui d'objet de politique étrangère. Les décisions qui déterminent aujourd'hui la capacité d'un pays à employer ces technologies ne sont plus prises dans des laboratoires mais dans des bureaux de contrôle des exportations, des salles de conseil d'administration et des files d'attente de raccordement électrique.

Trois faits récents rendent la question inévitable pour la France. Le 8 septembre 2026, Mistral AI a annoncé une levée de 3 milliards d'euros menée par un industriel coréen, portant sa valorisation au-delà de 21 milliards d'euros. À l'été 2026, les principales plateformes de distribution de modèles ont établi que les laboratoires chinois dominaient désormais l'écosystème des poids ouverts. Et, symétriquement, plusieurs acteurs américains — puis, depuis août 2026, certains laboratoires chinois — ont commencé à refermer ce qu'ils avaient ouvert.

Ces trois faits ne se lisent pas séparément. Ils décrivent une même reconfiguration : la souveraineté technologique n'est plus une question de propriété nationale d'un actif, mais une question de position dans une chaîne de dépendances multiples, dont chaque maillon peut être actionné par un acteur différent.

I.2 La question de recherche

Le hors-série répond à une question unique : **de quoi la France et l'Europe disposent-elles réellement en matière d'intelligence artificielle, et que perdraient-elles si une dépendance était interrompue ?**

Cette formulation écarte volontairement deux manières usuelles de poser le problème. Elle n'oppose pas « souveraineté » et « ouverture », car l'analyse montre que ces deux termes ne sont pas antagonistes mais situés sur des couches différentes. Elle n'oppose pas non plus « régulation » et « innovation », car les données disponibles ne permettent pas d'établir que la première explique le retard européen sur la seconde.

I.3 Le découpage retenu : six couches

Nexus retient une décomposition de la souveraineté en six couches ordonnées de l'amont vers l'aval : (1) les accélérateurs de calcul ; (2) les centres de données et l'énergie ; (3) le cloud ; (4) les modèles et leurs licences ; (5) les données ; (6) les normes, le droit et les talents.

L'intérêt de ce découpage est opératoire : il permet de constater qu'une même entité peut être souveraine sur une couche et totalement dépendante sur une autre, et que les politiques publiques adressent rarement plus d'une couche à la fois. Il permet aussi d'éviter l'erreur de catégorie la plus fréquente du débat français, qui consiste à présenter un processeur généraliste européen comme une réponse à la dépendance aux accélérateurs d'entraînement — ce qu'il n'est pas.

I.4 La distinction structurante

Le rapport articule l'ensemble de son analyse autour d'une distinction que le débat public confond systématiquement.

La **souveraineté de conformité** est la capacité à garantir, par le droit et le contrat, que les données traitées ne sont pas accessibles à une autorité étrangère et que l'utilisateur conserve la maîtrise de sa configuration. Elle se mesure en clauses, en certifications, en localisations. Elle est atteignable rapidement, à coût modéré, et protège contre l'accès non autorisé.

La **souveraineté de capacité** est la possibilité de concevoir, d'entraîner et de faire fonctionner un système sans l'autorisation d'un tiers. Elle se mesure en accélérateurs, en gigawatts, en équipes capables de reproduire un entraînement. Elle exige des décennies et des dizaines de milliards d'euros, et seule elle protège contre l'interruption.

L'Europe progresse rapidement sur la première et lentement sur la seconde. Cette asymétrie est le fait central du hors-série.

II. Cartographie du débat

II.1 Quatre positions

La position normative. La souveraineté passe par le droit : le règlement sur l'IA, le RGPD et le futur règlement sur le cloud et l'IA constituent un avantage comparatif européen. Objection : l'effet normatif suppose un marché assez attractif pour que l'on accepte d'en payer le prix de conformité, et il n'a jamais produit de champion industriel.

La position industrialiste. La souveraineté passe par le capital et l'infrastructure : financer des champions, construire des gigafactories, réserver une part de la commande publique. C'est la position de la commission d'enquête de l'Assemblée nationale et du rapport « Buy European Tech Act » de la chaire numérique de Sciences Po. Objection : le coût est considérable et la fenêtre étroite.

La position dérégulatrice. La souveraineté passe par la vitesse. Arthur Mensch, devant la commission d'enquête le 12 mai 2026 : « Réguler pour se défendre, en général, ça ne marche pas. Cela favorise les plus gros acteurs. » C'est la logique de l'omnibus numérique, adopté par le Parlement européen le 16 juin 2026. Objection, formulée par le CEPD et le Contrôleur européen dans leur avis conjoint du 11 février 2026 : la simplification ne doit pas se payer en droits fondamentaux.

La position des communs. La souveraineté passe par l'ouverture : des poids publiés, des données documentées, des évaluations reproductibles. C'est la position de la fondation Current AI, dotée de 400 millions et lancée au sommet de Paris en février 2025, et celle qu'a portée la déclaration de New Delhi de février 2026. Objection : l'ouverture ne suffit pas à produire l'indépendance, comme le montre le fait que les poids les plus employés soient désormais chinois.

Ces positions ne s'excluent pas : elles se hiérarchisent différemment selon l'horizon. La norme protège à court terme, l'ouverture permet un rattrapage à moyen terme, l'infrastructure seule produit une capacité à long terme.

II.2 Les acteurs

Cinq familles d'acteurs structurent le champ.

Les fournisseurs de couche matérielle : Nvidia (accélérateurs et, depuis septembre 2026, plateforme de distribution), AMD, Samsung et SK Hynix (mémoire), TSMC (fonderie), ASML (lithographie), Huawei et SMIC (filiale chinoise), SiPearl et VSORA (tentatives européennes).

Les laboratoires de frontière fermés : OpenAI, Anthropic, Google DeepMind et, depuis avril 2026, Meta Superintelligence Labs.

Les laboratoires à poids ouverts : Alibaba (Qwen), DeepSeek, Moonshot AI (Kimi), Z.ai (GLM) et MiniMax côté chinois ; Mistral AI côté européen ; Google (Gemma) et, de façon épisodique, OpenAI (gpt-oss) côté américain.

Les États et institutions : Commission européenne (norme, InvestAI, règlement cloud et IA), EuroHPC, États-Unis (contrôles à l'exportation), Chine (politique industrielle et étiquetage), Corée du Sud, Inde, Japon, Émirats arabes unis et Arabie saoudite.

Les acheteurs et régulateurs publics français : DINUM, ministère des Armées, Éducation nationale, collectivités territoriales, CNIL, ANSSI, INESIA, Cour des comptes et inspections générales.

II.3 Le point aveugle commun

Toutes ces positions partagent un angle mort : aucune ne chiffre publiquement le coût complet d'une capacité d'entraînement souveraine. Aucune évaluation française ne donne l'ordre de grandeur du parc d'accélérateurs, de l'énergie, des équipes et du cycle de renouvellement nécessaires. En l'absence de ce chiffre, le débat oppose des principes à des principes.

III. État empirique

III.1 La géographie du calcul

Selon Epoch AI, les États-Unis concentraient en mai 2025 environ 75 % de la performance mondiale des grappes de calcul GPU, la Chine environ 15 %, le reste du monde environ 10 %. Le rapport AI Index 2026 de Stanford HAI, publié le 13 avril 2026, situe la capacité mondiale à 17,1 millions d'équivalents H100, en croissance d'un facteur trente depuis 2021, et la capacité électrique des centres de données dédiés à 29,6 gigawatts.

L'investissement privé américain dans l'IA a atteint 285,9 milliards de dollars, contre 12,4 milliards pour la Chine — rapport de 23,1 à relativiser, les fonds d'orientation publics chinois étant estimés à environ 184 milliards hors périmètre. L'écart de performance entre les meilleurs modèles américains et chinois s'est réduit à 2,7 %, contre 17,5 à 31,6 points en mai 2023.

La disponibilité des accélérateurs est devenue un instrument de politique étrangère américaine. Le 15 janvier 2026, le Bureau of Industry and Security a fait passer les H200 et MI325X de la présomption de refus à l'examen au cas par cas pour la Chine, dans la limite d'environ 850 000 équivalents H200, assortie d'un droit de douane de 25 % ; en février 2026, le Département d'État a suspendu les expéditions dans l'attente d'une revue de sécurité. Le Conseil des relations étrangères américain a qualifié cette politique de « stratégiquement incohérente et inapplicable ». Le problème, pour l'Europe, n'est pas que les États-Unis restreignent : c'est qu'ils décident.

III.2 Le capital : la levée Mistral de septembre 2026

La série D annoncée le 8 septembre 2026 porte sur 3 milliards d'euros à une valorisation post-monnaire supérieure à 21 milliards, portant le total levé au-delà de 7,5 milliards d'euros. Le tour est mené par Samsung Electronics, avec le Scaleup Europe Fund géré par EQT et PSG Equity comme co-chefs de file. Nouveaux entrants : Advent, des fonds gérés

par BlackRock et le Grand-Duché de Luxembourg. L'entreprise déclare opérer dans vingt pays, compter plus de 125 clients grands comptes et viser 1 gigawatt de calcul en Europe d'ici 2030.

La trajectoire de valorisation est sans précédent européen : environ 240 millions d'euros en juin 2023, 5,8 milliards en juin 2024, 11,7 milliards en septembre 2025, plus de 21 milliards en septembre 2026. Le tour de septembre 2025 avait vu ASML investir 1,3 milliard d'euros et acquérir environ 11 % du capital, devenant ainsi le premier actionnaire individuel, avec un siège au comité stratégique.

Le revenu récurrent annualisé dépassait 400 millions de dollars en février 2026, multiplié par vingt en un an, avec une cible d'un milliard à fin 2026. Devant l'Assemblée nationale, le 12 mai 2026, Arthur Mensch a indiqué que 70 % du chiffre d'affaires était réalisé hors de France et que la commande publique française représentait environ 10 % de l'activité.

Trois investisseurs stratégiques figurent au capital et sont ou peuvent devenir fournisseurs : ASML (lithographie), Nvidia (accélérateurs), Samsung (mémoire). Le montant investi par Samsung n'est pas public ; la presse avance environ 1 milliard d'euros, estimation non confirmée.

III.3 L'infrastructure financée

Le premier centre de données détenu par Mistral AI, à Bruyères-le-Châtel (Essonne), opéré par Eclairion, développe 44 mégawatts et environ 13 800 accélérateurs Nvidia GB300 ; mise en service au deuxième trimestre 2026. Un site d'inférence de 10 mégawatts a été annoncé aux Ulis pour le troisième trimestre 2026. Le campus de Fouju (Seine-et-Marne), coentreprise avec MGX, Bpifrance et Nvidia annoncée le 20 mai 2025, vise 1,4 gigawatt à partir de 2028, avec une extension vers un réseau de 3 gigawatts annoncée le 1er juin 2026.

L'architecture de ce projet résume la situation française : capital franco-émirati, technologie américaine, électricité française, exploitant d'ancrage français. Chacun apporte ce dont il dispose, et la France apporte le facteur le moins rare.

III.4 Le basculement de l'open weights

Le rapport « State of Open Models: Summer 2026 » de Hugging Face (14 août 2026) établit que la famille Qwen d'Alibaba a franchi les 3 milliards de téléchargements sur six mois, contre 418 millions pour Google et 227 millions pour Meta ; que le plafond mensuel de taille des modèles ouverts chinois s'est situé entre 754 et 2 780 milliards de paramètres, contre moins de 130 milliards pour les modèles américains cinq mois sur sept ; et que, parmi les modèles chinois de plus de 20 milliards de paramètres, 59 % sont sous licence Apache 2.0 et 22 % sous licence MIT.

Sur OpenRouter, les modèles chinois ont consommé 34 250 milliards de jetons sur la semaine du 3 au 9 août 2026, dépassant les modèles américains pour la quinzième semaine consécutive. DeepSeek V4, publié le 24 avril 2026 sous licence MIT, atteint 1 600 milliards de paramètres et est tarifé 1,74 dollar par million de jetons en entrée contre 5 dollars pour GPT-5.5. Kimi K3, publié le 16 juillet 2026 par Moonshot AI, atteint 2 800 milliards de paramètres.

III.5 Le repli de l'ouverture

Meta a publié en avril 2026 Muse Spark, modèle propriétaire sans poids téléchargeables, alors que les licences Llama 4 n'étaient déjà pas approuvées par l'Open Source Initiative. OpenAI a publié gpt-oss le 5 août 2025 sous licence Apache 2.0, sans successeur de base à la mi-2026. Ce repli contredit la politique publique américaine : le plan « America's AI Action Plan » du 23 juillet 2025 comporte une section intitulée « Encourage Open Source and Open Weight AI ».

Le mouvement n'est plus unilatéral. Le 18 août 2026, Z.ai a reporté de deux semaines la publication des poids de GLM-5.3 pour renforcer ses contrôles de sécurité, le modèle atteignant 84,5 % sur le référentiel CyberGym et ayant identifié 2 436 vulnérabilités sur 269 projets. Les poids ont été publiés le 25 août 2026 sous licence propriétaire « glm-5.3 » avec revue de sécurité obligatoire au-delà de 10 milliards de dollars de chiffre d'affaires. Kimi K2.5 et K3 imposent des seuils de 100 millions d'utilisateurs actifs mensuels et de 20 millions de dollars ; MiniMax M3 impose une autorisation écrite au-delà de 20 millions de dollars ; le modèle haut de gamme d'Alibaba, Qwen 3.8-Max, ne dispose d'aucun dépôt public.

III.6 Le rachat de Hugging Face

Le 3 septembre 2026, Nvidia a annoncé l'acquisition de Hugging Face pour 12,93 milliards de dollars. La plateforme, cofondée par des entrepreneurs français, avait été valorisée 4,5 milliards de dollars en 2023 et réalisait environ 150 millions de dollars de chiffre d'affaires annuel. L'opération doit passer un examen antitrust. L'infrastructure de distribution des modèles ouverts — l'endroit où l'Europe s'approvisionne — passe donc sous le contrôle du fournisseur dominant des accélérateurs.

III.7 Le cadre européen

Le règlement (UE) 2024/1689 applique depuis le 2 août 2026 l'article 50 relatif à la transparence, la compétence des autorités nationales, le régime de sanctions (jusqu'à 15 millions d'euros ou 3 % du chiffre d'affaires mondial) et les bacs à sable. L'omnibus numérique, adopté par

le Parlement le 16 juin 2026 et par le Conseil le 29 juin 2026, a reporté les obligations relatives aux systèmes à haut risque de l'annexe III au 2 décembre 2027 et de l'annexe I au 2 août 2028.

L'exemption dont bénéficient les modèles ouverts (article 53, paragraphe 2) est étroite : elle dispense de la documentation technique et de l'information des fournisseurs en aval, sous quatre conditions cumulatives, mais ne s'applique pas aux modèles à risque systémique et ne lève ni l'obligation relative au droit d'auteur ni celle du résumé public des données d'entraînement. À mesure que les licences se durcissent, cette exemption couvre une part décroissante du marché réel.

La proposition de règlement sur le cloud et l'IA, présentée le 3 juin 2026, institue quatre niveaux d'assurance de souveraineté. Selon les estimations disponibles, environ 70 % des marchés relèveraient du niveau 1 — accessible aux filiales européennes des fournisseurs américains —, environ 20 % du niveau 2, moins de 10 % du niveau 3 et environ 1 % du niveau 4. Trois fournisseurs américains détiennent plus de 70 % du marché européen du cloud ; les fournisseurs européens environ 15 %.

III.8 La demande française

Selon l'Insee (21 juillet 2026), 18 % des entreprises implantées en France utilisaient l'IA en 2025, contre 10 % en 2024, avec 15 % pour les entreprises de 10 à 49 salariés et 58 % au-delà de 250 salariés, pour une moyenne de l'Union à 20 %. Selon Bpifrance (mars 2026), 47 % des PME ont lancé au moins un projet, contre 95 % des entreprises de plus de 200 salariés ; le principal frein déclaré est le recrutement (60 %).

La mission IGF-IGA-IGAS (avril 2026) recense 141 produits d'IA dans 80 entités publiques, estime à environ 40 % la part des agents utilisant l'IA de manière informelle, et relève que 90 % des métropoles sont engagées contre 73 % des communes de moins de 3 500 habitants sans aucun projet. Le baromètre de l'Observatoire Data Publica (14 novembre 2025) établit que 84 % des collectivités de plus de 3 500 habitants utilisent l'IA générative, contre 52 % en 2024, et que 82 % placent la localisation des données en France ou en Europe parmi leurs critères.

RTE recensait en mai 2026 environ 18 gigawatts de capacité réservée au titre d'environ 80 projets de centres de données, contre environ 5 gigawatts pour 40 projets fin 2024, pour une consommation actuelle d'environ 10 térawattheures, soit 2 % de la consommation nationale.

IV. Controverses

IV.1 L'open weights renforce-t-il ou affaiblit-il la souveraineté ?

Thèse de l'ouverture. Les poids ouverts permettent de relocaliser l'inférence, d'auditer partiellement le système, d'éviter l'enfermement propriétaire et de continuer à fonctionner si un fournisseur interrompt son service. Ils constituent le seul levier de rattrapage accessible à l'Europe à coût raisonnable.

Thèse du cheval de Troie. L'Institut Montaigne, sous la plume de François Godement, a qualifié le 7 septembre 2026 la politique chinoise de « dumping technologique », identifiant des risques de vulnérabilités intentionnelles, de comportements déclenchés, d'opacité des données d'entraînement et d'asymétrie entre une stratégie d'État coordonnée et une concurrence occidentale décentralisée.

État du débat. Les deux thèses sont partiellement vraies et portent sur des objets différents. L'ouverture des poids résout un problème de réversibilité ; elle ne résout pas un problème de caractérisation. On ne peut pas, à partir des seuls poids, déterminer ce sur quoi un modèle a été entraîné ni exclure un comportement conditionnel rare. Aucun incident de sécurité imputable à un modèle ouvert chinois n'a toutefois été documenté publiquement à grande échelle, et aucune évaluation publique française de ces modèles n'est disponible. La controverse restera ouverte tant que l'INESIA ne publiera pas d'évaluation.

IV.2 Le cas Mistral-GLM

Les 11 et 14 août 2026, Mistral AI a annoncé héberger sur son infrastructure, sans modification, le modèle chinois GLM-5.2 de Z.ai, avec les mêmes contrôles régionaux et engagements de service que ses propres modèles. Le directeur technique de l'entreprise l'a justifié : le modèle est à poids ouverts, il n'y avait donc « aucune bonne raison de ne pas le faire ».

Du point de vue de la conformité, la démonstration est imparable : le modèle tourne en Europe, les données ne sortent pas, le client contrôle. Du point de vue de la capacité, la conclusion est inverse : le champion européen distribue l'intelligence d'un laboratoire chinois parce qu'elle est meilleure ou moins chère sur certains usages. Ce cas est le révélateur le plus net de la distinction entre les deux souverainetés, et il explique pourquoi l'annonce a été lue par certains comme une trahison et par d'autres comme une preuve de maturité.

IV.3 La régulation est-elle un handicap ou un avantage ?

Thèse du handicap. Vingt-sept autorités nationales de mise en œuvre, un calendrier révisé deux fois, un cadre jugé flou par 52 % des PME interrogées par Bpifrance : la charge de conformité pèserait proportionnellement plus sur les petits acteurs européens que sur les grands acteurs américains.

Thèse de l'avantage. Le règlement crée une exigence que tout entrant doit satisfaire, et donne à l'Europe un instrument de négociation. Le CEPD et le Contrôleur européen rappellent que « la simplification est essentielle [...] mais pas au détriment des droits fondamentaux ».

État du débat. Les données disponibles ne permettent pas de trancher. Aucune étude n'établit de lien causal entre le cadre réglementaire européen et l'écart de capacité de calcul, qui préexistait largement au règlement. Un point fait en revanche consensus dans les enquêtes d'entreprises : c'est l'instabilité du calendrier, davantage que le niveau d'exigence, qui pèse sur l'investissement. De ce point de vue, l'omnibus numérique, qui reporte sans clarifier, est ambivalent.

IV.4 Faut-il une action spécifique de l'État au capital ?

Pour. La commission d'enquête de l'Assemblée nationale a recommandé le 15 juillet 2026 une participation de 5 % assortie de droits spécifiques permettant de bloquer l'entrée d'actionnaires étrangers sans faire obstacle aux levées de fonds, sur le fondement de l'article 31-1 de l'ordonnance du 20 août 2014.

Contre. Un tel dispositif peut dissuader les investisseurs internationaux dont l'entreprise a besoin pour financer son infrastructure, dans un secteur où la levée de fonds est la condition de survie. Il pose également une question de compatibilité avec le droit de l'Union sur la libre circulation des capitaux.

État du débat. La recommandation n'a pas été mise en œuvre ni publiquement instruite. L'absence d'arbitrage motivé est en soi une décision, et la moins informative.

****IV.5 La souveraineté énergétique est-elle un atout ou une rente**

convoitée ?**

La France dispose d'un surplus électrique évalué à 9 gigawatts et d'une électricité très peu carbonée. Deux lectures s'opposent : celle qui y voit un avantage à valoriser pour attirer les capacités de calcul, et celle qui y voit une ressource publique rare, susceptible d'être réservée par des acteurs extra-européens à travers des demandes de raccordement. Les

18 gigawatts de demandes recensées par RTE en mai 2026 ne préjugent pas de la consommation réelle — les installations existantes ne consomment en moyenne que 20 % de la puissance demandée — mais constituent une prise d’option sur le réseau. Les critères d’attribution de ces capacités n’ont pas fait l’objet d’un débat public à la hauteur de l’enjeu.

V. Angles morts

Le coût d’une capacité souveraine n’est pas chiffré. Aucune évaluation publique française ne donne le coût complet d’une capacité d’entraînement souveraine, renouvellement compris. C’est l’angle mort principal.

Le scénario d’interruption n’est pas documenté. La question posée par le directeur général de l’ANSSI devant la commission d’enquête — que se passe-t-il si l’accès aux technologies américaines est interrompu six mois ? — n’a pas reçu de réponse publique chiffrée.

L’articulation avec la trajectoire climatique est absente. Les projections de consommation électrique des centres de données (23 à 28 térawattheures en 2035 selon RTE) ne sont pas explicitement articulées avec la Stratégie nationale bas carbone.

La sécurité des modèles ouverts n’est pas évaluée publiquement en France. L’INESIA a vocation à combler ce vide ; aucune évaluation n’est disponible à ce jour.

La couche logicielle intermédiaire est ignorée. Le débat porte sur les accélérateurs et les modèles, et néglige les bibliothèques, les outils d’orchestration et les plateformes de distribution — dont le rachat de Hugging Face par Nvidia rappelle qu’elles sont des points de contrôle.

Les données d’usage réel manquent. Toutes les mesures de diffusion reposent sur des plateformes de développeurs. La part des modèles chinois, américains et européens dans les systèmes de production des entreprises et administrations européennes n’est mesurée par personne.

L’effet sur l’emploi n’est pas documenté en France. L’AI Index 2026 relève une baisse de 20 % du nombre de développeurs de 22 à 25 ans aux États-Unis depuis 2024 ; aucune étude française équivalente n’est publiée.

VI. Les neuf thèses falsifiables — rappel structurel

Le Rapport v2-0 énonce neuf thèses falsifiables numérotées THESE-HS-2026-010 à THESE-HS-2026-018, chacune assortie d'un énoncé, d'appuis empiriques datés, d'un niveau de robustesse et de conditions de réfutation observables.

THESE-HS-2026-010 — La souveraineté en IA se décompose en six couches, et l'Europe n'en maîtrise réellement qu'une. *Thèse consolidée.*

THESE-HS-2026-011 — La levée de septembre 2026 consolide un champion industriel sans réduire la dépendance matérielle. *Thèse consolidée.*

THESE-HS-2026-012 — Le capital du champion européen est devenu majoritairement extra-français, sans instrument de contrôle correspondant. *Hypothèse de travail.*

THESE-HS-2026-013 — L'écosystème des modèles à poids ouverts a basculé vers la Chine. *Thèse consolidée.*

THESE-HS-2026-014 — Le repli américain de l'ouverture est un choix économique, non une contrainte technique ou sécuritaire établie. *Hypothèse de travail.*

THESE-HS-2026-015 — L'ouverture chinoise se referme elle aussi, par le haut de gamme et par les licences. *Hypothèse de travail.*

THESE-HS-2026-016 — L'open weights est le seul levier de rattrapage européen à court terme, mais il déplace la dépendance au lieu de la supprimer. *Hypothèse de travail.*

THESE-HS-2026-017 — La souveraineté européenne a été redéfinie en conformité juridique et en localisation des données. *Thèse consolidée.*

THESE-HS-2026-018 — Le maillon faible français n'est pas le modèle mais la diffusion dans les territoires. *Hypothèse de travail.*

Bilan : quatre thèses consolidées sur convergence de sources indépendantes, cinq hypothèses de travail à consolider sur la période 2026-2030. Les hypothèses portent principalement sur les intentions (motivations des acteurs privés), sur les tendances récentes (fermeture chinoise observée depuis l'été 2026) et sur les liens de causalité (entre écart d'adoption et perte de souveraineté).

VII. Cahier des sources

VII.1 Hiérarchie appliquée

Nexus applique une hiérarchie stricte : sources primaires officielles, puis sources académiques et instituts de recherche, puis presse spécialisée, puis presse générale. Tout fait postérieur à janvier 2025 a été vérifié sur au moins deux sources indépendantes. Les faits établis (documents officiels, communiqués d'entreprise), les estimations (calculs d'analystes, parts de marché) et les annonces (engagements non exécutés) sont explicitement distingués dans le Rapport v2-O et dans le document de Recherche vO-I.

VII.2 Sources primaires mobilisées

Assemblée nationale (commission d'enquête sur les dépendances numériques) ; Commission européenne (plan « AI Continent », stratégies « Apply AI » et « AI in Science », proposition de règlement cloud et IA) ; EuroHPC (appel gigafactories, sélection des usines d'IA) ; Parlement et Conseil de l'Union (règlement 2024/1689 et omnibus numérique) ; CEPD et Contrôleur européen de la protection des données ; Cour des comptes ; mission IGF-IGA-IGAS ; DINUM ; CNIL ; Sénat ; ministère de l'Éducation nationale ; Insee ; RTE ; Bureau of Industry and Security ; Maison-Blanche ; ministère indien des Affaires étrangères. Publications d'entreprises : Mistral AI, MGX, Nvidia, Hugging Face, OpenAI, Z.ai, Moonshot AI, DeepSeek.

VII.3 Sources académiques et instituts

Stanford HAI (AI Index Report 2026, 13 avril 2026) ; Epoch AI (répartition de la performance des grappes, mai 2025) ; Institut Montaigne (7 septembre 2026) ; chaire numérique, gouvernance et souveraineté de Sciences Po (rapport « Buy European Tech Act », 2024) ; US-China Economic and Security Review Commission (mars 2026) ; Lawfare (analyse du règlement cloud et IA, juin 2026).

VII.4 Faits classés comme estimations

Les éléments suivants sont traités comme des estimations et non comme des faits établis : le montant investi par Samsung Electronics (environ 1 milliard d'euros, non confirmé) ; la répartition du capital de Mistral AI (35,5 % pour les cofondateurs, « un peu plus de 62 » % pour les investisseurs européens avant le dernier tour) ; les parts d'usage des modèles chinois sur OpenRouter (de 30 % à plus de 60 % selon les sources) ; les volumes et le chiffre d'affaires des accélérateurs Huawei Ascend ; la répartition des marchés par niveau d'assurance dans le futur règlement cloud et IA ; les valorisations privées d'OpenAI et d'Anthropic.

VII.5 Faits classés comme annonces non exécutées

Les gigafactories européennes (sélection prévue début 2027, mise en service dans les dix-huit mois suivants) ; le campus de Fouju (exploitation annoncée à partir de 2028) ; le réseau de 3 gigawatts de Campus AI ; l'objectif de 1 gigawatt de Mistral AI à 2030 ; l'outil souverain d'assistance aux enseignants attendu pour l'année scolaire 2026-2027 ; l'action spécifique recommandée par la commission d'enquête, non mise en œuvre ; le règlement cloud et IA, à l'état de proposition ; la migration du Health Data Hub, dont l'attributaire n'était pas confirmé publiquement à la date de rédaction.

VII.6 Points à vérifier avant publication

Trois points appellent une vérification directe : le statut exact de licence de plusieurs modèles Mistral (Medium 3.5, Codestral, Magistral Medium, Pixtral), sur lequel les sources secondaires divergent ; la participation précise de Samsung au capital, qui pourrait être publiée dans les documents réglementaires coréens ; et l'attributaire final de l'hébergement du Health Data Hub.

PARTIE II

Rapport de fond et thèses Nexus

RAPPORT DE FOND

Intelligence artificielle et souveraineté

Capital, calcul, modèles ouverts : ce que la France et l'Europe contrôlent réellement

Domaine : Hors-série

Périmètre : six couches de dépendance — accélérateurs de calcul, centres de données et énergie, cloud, modèles et licences, données, normes et talents

Comparaisons internationales : États-Unis, Chine, Corée du Sud, Inde, Émirats arabes unis et Arabie saoudite, Union européenne

Version 2-0 — Densifiée — Validation comité de lecture · 9 thèses : THESE-HS-2026-010 à 018

Résumé exécutif

Constat

La souveraineté en intelligence artificielle n'est pas un état binaire : elle se décompose en six couches — accélérateurs de calcul, centres de données et énergie, cloud, modèles et licences, données, normes et talents. L'Union européenne occupe une position dominante sur une seule, la norme, et une position de force sur un maillon amont, la lithographie avec ASML. Elle est marginale sur le calcul : les États-Unis concentraient en mai 2025 environ 75 % de la performance mondiale des grappes GPU, la Chine environ 15 %, le reste du monde environ 10 % (Epoch AI). Trois hyperscalers américains détiennent plus de 70 % du marché européen du cloud.

Le capital

Le 8 septembre 2026, Mistral AI a levé 3 milliards d'euros à une valorisation supérieure à 21 milliards, tour mené par Samsung Electronics avec le Scaleup Europe Fund (EQT) et PSG Equity — la plus importante levée en fonds propres d'une entreprise technologique européenne, portant le total levé au-delà de 7,5 milliards. Le premier actionnaire individuel reste ASML, entré à hauteur d'environ 11 % en septembre 2025. Trois investisseurs stratégiques sont ou peuvent devenir fournisseurs : ASML (lithographie), Nvidia (accélérateurs), Samsung (mémoire). La commission d'enquête de l'Assemblée nationale a recommandé le 15 juillet 2026 une action spécifique de l'État de 5 % : recommandation non mise en œuvre.

L'infrastructure

Le centre de données de Bruyères-le-Châtel (Essonne) développe 44 mégawatts sur environ 13 800 accélérateurs Nvidia GB300. Le campus de Fouju (Seine-et-Marne) vise 1,4 gigawatt à partir de 2028, avec Bpifrance, Nvidia et le fonds émirati MGX, et une extension annoncée vers 3 gigawatts. RTE recensait en mai 2026 environ 18 gigawatts de demandes de raccordement pour des centres de données, contre 5 gigawatts fin 2024. L'Union a lancé le 30 juillet 2026 son appel pour jusqu'à sept gigafactories, dépôt au 12 novembre 2026, mise en service dix-huit mois après une sélection attendue début 2027.

Les modèles ouverts

L'écosystème des poids ouverts a basculé vers la Chine. Selon Hugging Face (14 août 2026), la famille Qwen d'Alibaba a franchi les 3 milliards de téléchargements sur six mois, contre 418 millions pour Google et 227 millions pour Meta ; le plafond mensuel de taille des modèles chinois

s'est situé entre 754 et 2 780 milliards de paramètres, contre moins de 130 milliards pour les modèles américains cinq mois sur sept.

Symétriquement, Meta a publié en avril 2026 un modèle propriétaire sans poids téléchargeables, et OpenAI n'a pas donné de successeur à gpt-oss. Mistral AI demeure le seul acteur occidental publiant un modèle de plusieurs centaines de milliards de paramètres sous licence Apache 2.0.

Le retournement de l'été 2026

L'ouverture chinoise commence à se refermer. Le 18 août 2026, Z.ai a reporté de deux semaines la publication des poids de GLM-5.3 pour motif de cybersécurité, avant de les publier le 25 août sous licence propriétaire avec revue de sécurité obligatoire. Moonshot et MiniMax imposent des seuils commerciaux ; le modèle haut de gamme d'Alibaba n'a aucun dépôt public. Le 3 septembre 2026, Nvidia a annoncé le rachat de Hugging Face pour 12,93 milliards de dollars : la plateforme de distribution des modèles ouverts passe sous contrôle du fournisseur dominant des accélérateurs.

La distinction structurante

Nexus distingue la souveraineté de conformité — localisation des données, clauses, certification, atteignable vite et à coût modéré — et la souveraineté de capacité — concevoir, entraîner et faire fonctionner sans autorisation d'un tiers, qui exige des décennies. La première protège contre l'accès non autorisé ; seule la seconde protège contre l'interruption. Le règlement européen sur le cloud et l'IA, proposé le 3 juin 2026, réserverait selon les estimations moins de 10 % des marchés au niveau exigeant un fournisseur non contrôlé par un pays tiers.

Les territoires

18 % des entreprises implantées en France utilisaient l'IA en 2025 (Insee), contre 20 % dans l'Union : 15 % pour les entreprises de 10 à 49 salariés, 58 % au-delà de 250 salariés. 90 % des métropoles sont engagées contre 73 % des communes de moins de 3 500 habitants sans aucun projet, et environ 40 % des agents publics utilisent l'IA de manière informelle (IGF-IGA-IGAS, avril 2026). 82 % des collectivités placent pourtant la localisation des données parmi leurs critères. Une capacité inemployée n'est pas une capacité.

Méthode falsificationniste

Neuf thèses falsifiables (THESE-HS-2026-010 à 018) sont énoncées avec conditions de réfutation explicites : quatre consolidées sur convergence de sources indépendantes, cinq hypothèses de travail à

consolider sur 2026–2030. Révision selon le protocole de la charte Nexus dès lors qu’une condition de réfutation sera établie.

Trois scénarios à 2032

Consolidation européenne (gigafactories livrées, seuils de souveraineté significatifs, adoption rejoignant la moyenne européenne) ; conformité sans capacité (trajectoire tendancielle, droit respecté sans capacité propre) ; rupture d’approvisionnement (interruption des accélérateurs, écart accru de deux à trois générations en trois ans). Les mesures qui préservent la valeur dans le troisième scénario — poids ouverts, réversibilité, compétences internes, calcul mutualisé — sont aussi celles qui coûtent le moins dans les deux premiers. Ce sont des mesures sans regret.

Préambule

Le rapport hors-série Nexus « Intelligence artificielle et souveraineté » s’inscrit dans le corpus Nexus comme méta-rapport transversal. Il traite un objet — la capacité d’une nation à décider seule de l’usage, de la disponibilité et de l’évolution des systèmes d’intelligence artificielle qu’elle emploie — qui traverse l’ensemble des piliers du corpus : Économie (industrie, services, emploi), Société (santé, information, défense), Infrastructures (énergie, réseaux, calcul, cloud) et Éducation (formation, recherche, compétences).

Le sujet est devenu, en moins de quatre ans, un objet de politique étrangère. Ce qui se jouait en 2022 dans des laboratoires de recherche se joue désormais dans des décrets de contrôle des exportations, des accords de fourniture d’accélérateurs graphiques, des files d’attente de raccordement électrique et des tours de table capitalistiques. Trois séquences récentes obligent à reprendre la question à sa base. La première est la levée de fonds de 3 milliards d’euros annoncée par Mistral AI le 8 septembre 2026, menée par un industriel coréen. La deuxième est le basculement de l’écosystème des modèles à poids ouverts vers la Chine, désormais mesurable sur les principales plateformes de distribution. La troisième est le mouvement inverse observé chez plusieurs acteurs américains — et, depuis l’été 2026, chez certains laboratoires chinois : une fermeture progressive, par le haut de gamme et par les licences.

Méthodologie falsificationniste popperienne. Neuf thèses falsifiables sont énoncées avec leurs conditions de réfutation explicites, numérotées THESE-HS-2026-010 à THESE-HS-2026-018. Quatre sont consolidées sur la base de données empiriques convergentes et robustes ; cinq restent en hypothèse de travail nécessitant complément

empirique. Le rapport sera révisé selon le protocole de la charte Nexus en cas de données nouvelles pertinentes : publication des rapports institutionnels attendus, nouveaux tours de financement, évolutions du cadre réglementaire européen, modification des politiques de licence des principaux laboratoires.

Le ton retenu est analytique. Ce rapport ne prescrit pas une doctrine industrielle ; il établit ce qui est documenté, distingue ce qui est estimé de ce qui est annoncé, et énonce à quelles conditions ses propres conclusions devraient être abandonnées.

I. Cadrage : ce que « souveraineté » veut dire en intelligence artificielle

I.1 Une notion contestée

Le mot « souveraineté » est employé, dans le débat sur l'intelligence artificielle, dans au moins quatre acceptions incompatibles.

La première est juridictionnelle : un système est souverain si les données qu'il traite ne sont pas accessibles à une autorité étrangère. C'est la définition retenue par la qualification SecNumCloud de l'ANSSI, qui exige un siège social dans l'Union, un hébergement dans l'Union et une participation individuelle d'actionnaire non européen plafonnée à 24 %.

La deuxième est contractuelle : un système est souverain si son utilisateur conserve la maîtrise de ses données, de la configuration du modèle et des conditions de sortie. C'est la définition que retient Mistral AI, qui décrit la souveraineté comme le contrôle sur quatre dimensions — les données qui restent dans le périmètre de l'organisation, des modèles contrôlables et personnalisables, un calcul privé et prévisible, des systèmes en production pleinement contrôlables et auditaibles.

La troisième est capacitaire : un système est souverain si la nation qui l'emploie peut le concevoir, l'entraîner et le faire fonctionner sans l'autorisation d'un tiers. C'est la définition implicite du plan américain « Winning the Race » de juillet 2025, et celle que résumait Michael Kratsios au sommet de New Delhi en février 2026 : « La véritable souveraineté en IA, c'est posséder la meilleure technologie de sa catégorie. »

La quatrième est commerciale : « IA souveraine » désigne un segment de marché. Nvidia a déclaré 30 milliards de dollars de chiffre d'affaires sur ce segment pour son exercice 2026, triplé en un an — soit environ 1 % de ses ventes totales. Dans cette acception, la souveraineté est ce que l'on achète à un fournisseur unique.

Ces acceptions ne sont pas équivalentes. Un système peut satisfaire la première et la deuxième tout en échouant complètement sur la troisième. C'est précisément la situation européenne, et c'est le point aveugle de la plupart des annonces publiques.

I.2 Les six couches de la dépendance

Nexus retient une décomposition en six couches, ordonnées de l'amont vers l'aval.

Couche 1 — Les accélérateurs de calcul. Circuits spécialisés pour l'entraînement et l'inférence. Nvidia occupe une position quasi monopolistique sur le segment de l'entraînement de modèles de fondation. Il n'existe aucun accélérateur européen d'entraînement en production. Le processeur Rhea1 de SiPearl, sorti des chaînes de TSMC et en phase de validation en mai 2026, est un processeur généraliste doté de 80 cœurs Arm Neoverse V1 et de 64 gigaoctets de mémoire à haute bande passante : utile pour le calcul intensif, il ne constitue pas une alternative aux accélérateurs pour l'entraînement de modèles de fondation. Cette distinction est systématiquement brouillée dans la communication publique.

Couche 2 — Les centres de données et l'énergie. La contrainte n'est plus le foncier mais le raccordement électrique. En France, environ 18 gigawatts de capacité étaient réservés en mai 2026 au titre d'environ 80 projets, contre environ 5 gigawatts pour 40 projets fin 2024 (RTE). La consommation actuelle des centres de données, environ 10 térawattheures, représente environ 2 % de la consommation nationale ; RTE projette 15 à 20 térawattheures en 2030 et 23 à 28 en 2035. L'électricité décarbonée et disponible est le principal atout français sur cette couche — et un atout convoité.

Couche 3 — Le cloud. Trois fournisseurs américains détiennent plus de 70 % du marché européen ; les fournisseurs européens environ 15 %. La commission d'enquête de l'Assemblée nationale a avancé le chiffre de 83 % des dépenses cloud européennes captées par les acteurs américains, soit environ 264 milliards d'euros par an.

Couche 4 — Les modèles et leurs licences. C'est la couche la plus mobile, et celle sur laquelle une position peut être reconquise en quelques mois. C'est aussi celle où l'Europe dispose d'un acteur de premier plan mondial.

Couche 5 — Les données. Données d'entraînement, données d'usage, données métier. L'enjeu se déplace du stockage vers la traçabilité : le règlement sur l'IA impose un résumé public des données d'entraînement, obligation à laquelle même les modèles ouverts ne peuvent se soustraire.

Couche 6 — Les normes, le droit et les talents. L'Union européenne est ici dominante par sa capacité normative. Le rapport AI Index 2026 relève à l'inverse que le flux net de chercheurs en IA vers les États-Unis a chuté de 89 % depuis 2017 — indicateur ambigu, qui peut signaler autant une saturation américaine qu'un rééquilibrage.

I.3 Souveraineté, autonomie stratégique, résilience

Trois notions voisines méritent d'être distinguées. La *souveraineté* est la capacité de décider. L'*autonomie stratégique* est la capacité d'agir sans dépendre d'un tiers sur les fonctions jugées critiques — notion plus étroite et plus réaliste. La *résilience* est la capacité de continuer à fonctionner en mode dégradé si une dépendance est interrompue.

La formulation la plus opérationnelle du problème a été donnée par Vincent Strubel, directeur général de l'ANSSI, devant la commission d'enquête de l'Assemblée nationale : « Si demain, pendant six mois, on est privés d'accès à la technologie américaine, privés des mises à jour des logiciels américains, on a un énorme problème. » Cette phrase ne décrit pas un défaut de souveraineté : elle décrit un défaut de résilience. Ce sont deux chantiers différents, de coûts et de calendriers différents.

II. État empirique : la géographie du calcul

II.1 La concentration américaine

Selon Epoch AI, les États-Unis concentraient en mai 2025 environ 75 % de la performance mondiale des grappes de calcul GPU, la Chine environ 15 %, le reste du monde — Allemagne, Japon et France compris — environ 10 %. Epoch observe que les puissances traditionnelles du calcul intensif « ne jouent plus qu'un rôle marginal » dans le paysage des grappes d'intelligence artificielle : le basculement du calcul académique vers le calcul d'entreprise s'est accompagné d'un basculement géographique.

Le rapport AI Index 2026 de Stanford HAI, publié le 13 avril 2026, donne les ordres de grandeur globaux : 171 millions d'équivalents H100 de capacité mondiale, soit une multiplication par trente depuis 2021 ; 29,6 gigawatts de capacité électrique de centres de données dédiés, soit l'équivalent de la pointe de consommation de New York ; 285,9 milliards de dollars d'investissement privé américain contre 12,4 milliards pour la Chine, soit un rapport de 23,1 — rapport qui doit être relativisé, les fonds d'orientation publics chinois étant estimés à environ 184 milliards de dollars hors périmètre.

II.2 Le rattrapage chinois

Le même rapport établit que l'écart de performance entre les meilleurs modèles américains et chinois s'est réduit à 2,7 %, contre 17,5 à 31,6 points en mai 2023. Ce chiffre est l'un des plus commentés de l'année 2026 : il signifie que la supériorité américaine ne se mesure plus en capacité de modèle, mais en capacité de calcul, en capitaux et en position de marché.

Le rattrapage s'opère sous contrainte matérielle. Les accélérateurs Huawei Ascend équiperont environ 45 % des centres de données chinois, pour un chiffre d'affaires annuel d'environ 15 milliards de dollars — estimations de presse spécialisée, non confirmées par source primaire. Leur fabrication dépend de SMIC, plafonné à une gravure de 7 nanomètres avec des rendements inférieurs à 50 %, faute d'accès aux équipements de lithographie extrême ultraviolet d'ASML.

La politique américaine de contrôle des exportations s'est elle-même assouplie de manière discutée. Le 15 janvier 2026, le Bureau of Industry and Security a fait passer les accélérateurs Nvidia H200 et AMD MI325X de la présomption de refus à l'examen au cas par cas pour la Chine, dans la limite d'environ 850 000 équivalents H200 et de 50 % des ventes américaines cumulées, assortie d'un droit de douane de 25 % institué par la proclamation du 14 janvier 2026. Les accélérateurs de classe Blackwell demeurent interdits. Le 21 janvier 2026, la commission des affaires étrangères de la Chambre des représentants a adopté par 42 voix contre 2 l'AI OVERWATCH Act, qui instituerait une interdiction statutaire de deux ans sur les Blackwell et un droit de veto du Congrès sur les licences individuelles. En février 2026, le Département d'État a suspendu les expéditions de H200 dans l'attente d'une revue de sécurité. Le Conseil des relations étrangères américain a qualifié cette politique de « stratégiquement incohérente et inapplicable ».

Il faut en tirer une conséquence pour l'Europe : la disponibilité des accélérateurs est devenue un instrument de politique étrangère américaine, mobilisé et démobilité en quelques semaines. Le problème n'est pas que les États-Unis restreignent : c'est qu'ils décident.

II.3 La position européenne

L'Europe construit. Le plan d'action « AI Continent », publié le 9 avril 2025, prévoit 19 usines d'intelligence artificielle et jusqu'à cinq gigafactories, mobilise 200 milliards d'euros au titre de l'initiative InvestAI dont 20 milliards pour les gigafactories, et vise un triplement de la capacité de centres de données de l'Union en cinq à sept ans. Le réseau des usines d'IA comptait treize sites après la deuxième vague de sélection de mars 2025, dont l'usine française AI2F pilotée par GENCI avec Inria, le CNRS, le CEA et le CINES ; les six projets de cette vague représentent environ 485 millions d'euros.

L'entreprise commune EuroHPC a lancé le 30 juillet 2026 l'appel à candidatures pour les gigafactories : jusqu'à sept installations, dépôt le 12 novembre 2026, sélection attendue début 2027, mise en service dans les dix-huit mois suivants, plus de 20 milliards d'euros d'investissement privé attendus. En France, le consortium AION mené par Scaleway avec OpCore, GENCI, Inria, SiPearl, Kyutai, Sopra Steria et d'autres, porte une candidature de 288 000 équivalents H100 pour 200 mégawatts.

Le calendrier est le point essentiel. Une gigafactory sélectionnée début 2027 et mise en service dix-huit mois plus tard entre en service à la mi-2028, à un moment où le parc mondial aura de nouveau été multiplié. L'Europe ne rattrape pas un retard : elle réduit un écart qui continue de croître en valeur absolue.

II.4 L'énergie comme facteur limitant

La France dispose d'un avantage structurel : une électricité abondante, pilotable et très peu carbonée. Arthur Mensch a chiffré à 9 gigawatts le surplus électrique français, en alertant sur le risque de le voir capté par des centres de données américains faute de décision européenne dans les deux ans.

RTE apporte une nuance décisive : les centres de données existants ne consomment en moyenne que 20 % de la puissance qu'ils ont demandée, la montée en charge s'étalant sur dix à quinze ans. Les 18 gigawatts de demandes de raccordement enregistrées en mai 2026 ne doivent donc pas être lus comme une consommation imminente. Ils doivent en revanche être lus comme une réservation de droits sur un réseau, c'est-à-dire comme une prise d'option stratégique. Qui obtient la file d'attente obtient le calcul.

II.5 Le cas français

La France cumule un atout énergétique, une position académique honorable, un champion industriel de niveau mondial et une absence totale de capacité de production d'accélérateurs. Elle abrite, avec le campus de Fouju, le plus grand projet européen annoncé — 1,4 gigawatt, porté par une coentreprise associant Bpifrance, Mistral AI, Nvidia et le fonds émirati MGX, avec une extension annoncée le 1er juin 2026 vers un réseau de 3 gigawatts.

L'architecture de ce projet résume la situation : le capital est franco-émirati, la technologie est américaine, l'électricité est française, l'exploitant d'ancrage est français. Chacun apporte ce dont il dispose. Nommer cela « souveraineté » est un abus de langage ; nommer cela « dépendance » l'est tout autant. Il s'agit d'une interdépendance asymétrique dans laquelle la France apporte le facteur le moins rare.

III. Le capital : la levée de fonds Mistral AI de septembre 2026

III.1 Les faits

Le 8 septembre 2026, Mistral AI a annoncé la clôture d'une série D de 3 milliards d'euros, portant sa valorisation post-monnaire au-delà de 21 milliards d'euros (environ 24 milliards de dollars). L'opération est présentée comme la plus importante levée en fonds propres jamais réalisée par une entreprise technologique européenne. Le total des fonds propres levés depuis la création dépasse 7,5 milliards d'euros.

Le tour est mené par Samsung Electronics, avec pour co-chefs de file le Scaleup Europe Fund géré par EQT et PSG Equity. Les nouveaux entrants sont Advent, des fonds gérés par BlackRock et le Grand-Duché de Luxembourg. Les investisseurs historiques participants comprennent a16z, ASML, Belfius, BNP Paribas CIB, Bpifrance, Carmignac, DST Global, Eurazeo, General Catalyst, Headline, Hillspire, Index Ventures, Korelya Capital, Lightspeed, NVIDIA, le fonds Solar de Phoenix Court et Salesforce Ventures.

L'entreprise déclare opérer dans vingt pays, compter plus de 125 clients grands comptes — dont Airbus, ASML et HSBC — et viser 1 gigawatt de capacité de calcul en Europe d'ici 2030. Selon TechCrunch, l'entrée de Samsung a reçu l'aval des autorités françaises.

III.2 Le fil des tours précédents

La trajectoire de valorisation est l'une des plus rapides de l'histoire industrielle européenne : environ 240 millions d'euros en juin 2023, plus de 2 milliards en décembre 2023, 5,8 milliards en juin 2024, 11,7 milliards en septembre 2025, plus de 21 milliards en septembre 2026. En trois ans et quatre mois, la valorisation a été multipliée par quatre-vingt-sept.

Le tour du 9 septembre 2025 mérite une attention particulière : ASML, fabricant néerlandais d'équipements de lithographie, y a investi 1,3 milliard d'euros sur les 1,7 milliard levés, acquérant environ 11 % du capital et devenant ainsi le premier actionnaire individuel de l'entreprise, devant chacun des trois cofondateurs. Roger Dassen, directeur financier d'ASML, siège au comité stratégique. La logique était double : industrielle — appliquer l'IA à l'optimisation des équipements de lithographie extrême ultraviolet — et stratégique, les deux sociétés affichant la volonté d'éviter une dépendance excessive aux acteurs américains ou chinois.

Un troisième financement, non capitalistique, complète le tableau : en mars 2026, un consortium de sept banques a apporté 830 millions de dollars de dette destinés au centre de données de l'Essonne. C'est la

première fois qu'un acteur européen de l'IA finance son infrastructure par de la dette bancaire plutôt que par de la dilution — indicateur de maturation industrielle.

III.3 L'entrée de Samsung

Le montant investi par Samsung Electronics n'a pas été rendu public. La presse économique avance un ordre de grandeur d'environ 1 milliard d'euros, soit approximativement un tiers du tour : cette estimation n'est pas confirmée. Le pourcentage de capital correspondant, l'existence d'un siège au comité stratégique et les éventuels engagements d'approvisionnement ne sont pas documentés publiquement.

La rationalité stratégique est en revanche explicite. Samsung cherche à élargir la base de clients de son activité de semi-conducteurs de mémoire et à capter la demande liée à la montée en charge des infrastructures d'IA. Samsung et SK Hynix contrôlent ensemble environ les deux tiers de la production mondiale de mémoire. Le groupe vise par ailleurs le marché des organisations qui conservent le contrôle direct de leurs données, de leurs modèles et de leur environnement d'exploitation — segment qui recoupe ses propres besoins industriels dans les semi-conducteurs et la fabrication avancée.

Cette entrée s'inscrit dans une politique nationale. Le 29 juin 2026, la Corée du Sud a annoncé un plan industriel décennal d'une ampleur exceptionnelle, dont un investissement conjoint de Samsung et SK Hynix d'environ 518 milliards de dollars dans des capacités de production de semi-conducteurs avancés. Le président Lee Jae Myung a formulé l'objectif : sécuriser les éléments fondamentaux de l'IA — semi-conducteurs, IA physique, centres de données, robotique — « plus vite que tout autre pays ».

Il faut relever que Samsung n'est pas le premier investisseur asiatique au capital de Mistral AI. Korelya Capital, fonds franco-coréen adossé au groupe Naver, figurait déjà parmi les investisseurs historiques. La nouveauté n'est pas l'origine du capital : c'est le fait qu'un fournisseur de composants prenne la tête d'un tour de table dans le champion européen des modèles.

III.4 La structure du capital et la question du contrôle

Trois investisseurs stratégiques figurent aujourd'hui au capital de Mistral AI, et tous trois sont ou peuvent devenir ses fournisseurs : ASML pour la lithographie, Nvidia pour les accélérateurs, Samsung pour la mémoire. Cette configuration n'est pas anormale dans une chaîne de valeur capitaliste ; elle appelle une vigilance sur les conditions commerciales et sur l'indépendance des choix d'architecture.

Sur la répartition du capital, l'enquête la plus détaillée disponible indique que les trois cofondateurs détiendraient ensemble environ 35,5 % du capital et que les investisseurs français et européens détenaient, avant le tour de septembre 2026, « un peu plus de 62 % ». Ces chiffres ne sont pas confirmés par l'entreprise et sont mécaniquement érodés à chaque tour. La cartographie des origines est large : américaine (Lightspeed, a16z, General Catalyst, NVIDIA, PSG Equity, BlackRock), asiatique et moyen-orientale (Samsung, DST Global, Sanabil, Mubadala), européenne (ASML, Bpifrance, Eurazeo, Carmignac, BNP Paribas), supranationale (Scaleup Europe Fund, Luxembourg).

Face à cette situation, la commission d'enquête de l'Assemblée nationale sur les dépendances structurelles et les vulnérabilités systémiques dans le secteur du numérique — créée le 3 février 2026, présidée par Philippe Latombe, rapportée par Cyrielle Chatelain, rapport adopté le 8 juillet 2026 et rendu public le 15 juillet 2026 après 45 auditions et 47 recommandations — a proposé l'institution d'une action spécifique (« golden share ») de l'État dans Mistral AI et ChapsVision : une participation de 5 % assortie de droits spécifiques permettant de bloquer l'entrée d'actionnaires étrangers sans faire obstacle aux levées de fonds. Le fondement invoqué est l'article 31-1 de l'ordonnance du 20 août 2014, qui permet à l'État des droits particuliers indépendamment de sa part de capital, dont un droit d'opposition aux cessions d'actifs stratégiques. La mise en œuvre supposerait une augmentation de la participation de Bpifrance. À la date de rédaction, cette recommandation n'a pas été mise en œuvre.

III.5 L'infrastructure financée

Le premier centre de données détenu par Mistral AI se trouve à Bruyères-le-Châtel (Essonne), opéré par Eclairion, d'une capacité de 44 mégawatts, équipé d'environ 13 800 accélérateurs Nvidia GB300, mis en service au deuxième trimestre 2026. Un second site d'inférence de 10 mégawatts, hébergé par Digital Realty aux Ulis, a été annoncé le 28 mai 2026 pour le troisième trimestre 2026.

Le projet Campus AI, annoncé le 20 mai 2025 entre MGX, Bpifrance, Mistral AI et NVIDIA, prévoit 1,4 gigawatt à Fouju (Seine-et-Marne) : travaux au second semestre 2026, exploitation à partir de 2028. Le 1er juin 2026, les partenaires ont annoncé l'extension vers un réseau de 3 gigawatts en France, Mistral AI étant locataire d'ancrage avec accès prioritaire au calcul. Mistral AI a par ailleurs conclu un accord pluriannuel d'infrastructure GPU de plusieurs milliards de dollars avec Microsoft.

Le point de tension est identifié par plusieurs analyses : l'entreprise qui incarne la souveraineté européenne est devenue l'un des principaux clients européens de Nvidia, et la souveraineté qu'elle offre est redéfinie comme résidence des données et conformité réglementaire plutôt que comme autodétermination technologique.

IV. Les modèles ouverts : un basculement vers la Chine

IV.1 Ce que mesurent les plateformes

Le rapport « State of Open Models: Summer 2026 » publié par Hugging Face le 14 août 2026 est la source la plus complète disponible. La famille Qwen d'Alibaba a franchi les 3 milliards de téléchargements sur six mois, contre 418 millions pour les modèles ouverts de Google et 227 millions pour ceux de Meta sur l'année 2026. Le plafond mensuel de taille des modèles ouverts publiés par des laboratoires chinois s'est situé entre 754 et 2 780 milliards de paramètres ; celui des laboratoires américains est resté sous 130 milliards cinq mois sur sept.

Une analyse indépendante d'avril 2026 établissait que Qwen représentait plus de 50 % des téléchargements mondiaux de modèles ouverts en mars 2026, avec près d'un milliard de téléchargements cumulés et 153,6 millions pour le seul mois de février. Sur la plateforme d'agrégation OpenRouter, les modèles chinois ont consommé 34 250 milliards de jetons sur la semaine du 3 au 9 août 2026, dépassant les modèles américains pour la quinzième semaine consécutive.

IV.2 Les laboratoires chinois

DeepSeek a publié le 24 avril 2026 les modèles V4-Pro (1 600 milliards de paramètres, 49 milliards actifs) et V4-Flash (284 milliards, 13 milliards actifs) sous licence MIT, avec une fenêtre de contexte d'un million de jetons, tarifés 1,74 et 3,48 dollars par million de jetons contre 5 et 30 pour GPT-5.5 à la même date. DeepSeek affirme n'accuser qu'un retard de trois à six mois sur les meilleurs modèles fermés.

Moonshot AI a publié Kimi K3 le 16 juillet 2026, avec ouverture des poids le 27 juillet : 2 800 milliards de paramètres, soit le plus grand modèle à poids ouverts publié à cette date, premier sur le classement Arena.ai Frontend Code devant Claude Fable 5 et GPT-5.6 Sol.

Alibaba, Z.ai (anciennement Zhipu), MiniMax, Tencent et l'Académie de Pékin pour l'intelligence artificielle complètent le paysage. Selon Martin Casado (Andreessen Horowitz), cité par MIT Technology Review, il y a « 80 % de chances » qu'une jeune pousse de la Silicon Valley utilisant une pile open source fasse tourner un modèle chinois.

IV.3 Les ressorts de la stratégie chinoise

Quatre ressorts se combinent.

Le premier est industriel : ouvrir les poids permet de diffuser une architecture, donc d'imposer des conventions techniques — formats de quantification, formats d'appel d'outils, tokenisation — que l'écosystème adopte par défaut. C'est la logique des normes, appliquée à l'IA.

Le deuxième est commercial : l'ouverture est un canal d'acquisition. Le modèle est gratuit, l'inférence hébergée est payante, et l'entreprise capte la clientèle qui a d'abord expérimenté gratuitement.

Le troisième est politique : le Conseil des affaires de l'État a publié en août 2024 un projet encourageant les universités à valoriser les contributions open source dans l'évaluation académique. Lors de la réunion ministérielle de l'APEC à Chengdu, dont la déclaration a été publiée le 26 juillet 2026, les vingt et une économies membres, États-Unis et Chine compris, ont affirmé soutenir « les modèles et projets open source qui mettent en œuvre une assurance de sécurité renforcée ».

Le quatrième est contraint : privés d'accès aux accélérateurs de dernière génération, les laboratoires chinois ont un intérêt objectif à maximiser la diffusion de leurs modèles, qui leur apporte des retours d'usage, des dérivés et une influence sans coût d'infrastructure supplémentaire.

L'Institut Montaigne, sous la plume de François Godement, a qualifié le 7 septembre 2026 cette politique de « dumping technologique », identifiant des risques de vulnérabilités intentionnelles, d'opacité des données d'entraînement et d'asymétrie entre une stratégie d'État coordonnée et une concurrence occidentale décentralisée. La commission américaine US-China Economic and Security Review a qualifié en mars 2026 cette position d'« avantage compétitif auto-renforçant ».

IV.4 Les limites de la mesure

Trois précautions s'imposent. D'abord, les téléchargements ne sont pas des déploiements : le rapport de Hugging Face relève que 85,6 % des dépôts comptent moins de 200 téléchargements cumulés et que 1,5 % des dépôts concentrent 99,2 % des téléchargements. Ensuite, les modèles de moins d'un milliard de paramètres représentent 83 % des téléchargements cumulés et les modèles de plus de 100 milliards seulement 1 % : la mesure est dominée par de petits modèles d'embarquement et de plongement lexical, sans rapport avec la frontière technologique. Enfin, les parts d'usage sur OpenRouter varient fortement selon les périodes et les périmètres : les publications de 2026 avancent des valeurs allant de 30 % à plus de 60 %. Aucune de ces mesures ne renseigne directement sur la part des modèles chinois dans les systèmes de production des grandes entreprises européennes — donnée qui, à notre connaissance, n'est pas publiée.

V. Le repli de l'ouverture

V.1 Meta

Meta a publié en avril 2026 Muse Spark, développé par Meta Superintelligence Labs : modèle propriétaire, accessible uniquement par interface applicative en nuage, sans poids téléchargeables. La société indique que les modèles Llama existants « resteront disponibles en open source », sans engagement sur les générations futures. Le cumul historique des téléchargements de Llama est estimé à 1,2 milliard. Andrew Ng a qualifié ce virage de « perte significative ».

Le repli était préparé par les licences. Llama 4 Scout et Maverick ne sont pas sous licence approuvée par l'Open Source Initiative : validation manuelle, seuil de 700 millions d'utilisateurs actifs mensuels déclenchant des obligations de marque et de dénomination des dérivés. Meta avait par ailleurs refusé, le 18 juillet 2025, de signer le code de bonnes pratiques européen pour les modèles à usage général, qu'OpenAI et Anthropic ont signé.

V.2 OpenAI

OpenAI a publié le 5 août 2025 gpt-oss-120b et gpt-oss-20b sous licence Apache 2.0, ses premiers modèles à poids ouverts depuis GPT-2 en 2019. Des variantes gpt-oss-safeguard ont suivi en octobre 2025. À la mi-2026, aucun modèle de base successeur n'avait été publié. Le geste a donc été ponctuel : une réponse tactique à la pression concurrentielle de DeepSeek, non l'inauguration d'une politique.

Il faut relever la contradiction avec le cadre politique américain : le plan « Winning the Race » publié par la Maison-Blanche le 23 juillet 2025 comporte une section explicitement intitulée « Encourage Open Source and Open Weight AI » et charge la NTIA de promouvoir l'adoption de modèles ouverts par les PME. La politique publique américaine encourage l'ouverture ; les entreprises américaines s'en écartent.

V.3 Mistral AI

Mistral AI a publié le 2 décembre 2025 la famille Mistral 3 sous licence Apache 2.0 : Mistral Large 3 (675 milliards de paramètres, 41 milliards actifs) et Ministral 3 en versions 14B, 8B et 3B. L'entreprise affirme que « l'avenir de l'IA est ouvert » et que l'ouverture met l'IA « entre les mains des gens par l'intelligence distribuée ».

Le portefeuille n'est cependant pas intégralement ouvert : plusieurs références — Mistral Medium 3.5, Codestral, Magistral Medium, certaines versions de Pixtral — relèvent de licences restrictives ou d'un accès par interface applicative uniquement. Les sources consultées divergent sur

le statut exact de certaines d'entre elles. Ce point doit être vérifié directement auprès de la documentation de l'éditeur avant toute affirmation publique.

Mistral AI demeure néanmoins, à ce jour, le seul acteur occidental à publier un modèle de plusieurs centaines de milliards de paramètres sous licence approuvée par l'Open Source Initiative.

V.4 Les signaux chinois

Le mouvement de fermeture n'est plus unilatéral. Le 18 août 2026, Z.ai a annoncé reporter de deux semaines la publication des poids de GLM-5.3, indiquant renforcer ses contrôles de sécurité avant mise à disposition large. Le modèle atteignait 84,5 % sur le référentiel CyberGym — devant des modèles fermés concurrents — et avait identifié 2 436 vulnérabilités sur 269 projets, dont 1 097 de sévérité moyenne à critique. C'est, à la connaissance des sources consultées, le premier cas documenté de retenue volontaire d'un laboratoire chinois pour motif de sécurité.

Les poids ont été publiés le 25 août 2026, mais sous une licence propriétaire « glm-5.3 » et non plus sous MIT : les opérateurs de modèle en tant que service dont le chiffre d'affaires cumulé sur douze mois glissants dépasse 10 milliards de dollars doivent obtenir une revue de sécurité de Z.ai avant usage commercial.

Un audit de licences publié le 16 août 2026 documente une diffusion plus générale des seuils commerciaux : Kimi K2.5 et K3 imposent des obligations de marque au-delà de 100 millions d'utilisateurs actifs mensuels ou 20 millions de dollars de revenu mensuel, et un accord séparé au-delà de 20 millions de dollars de revenu sur douze mois glissants ; MiniMax M3 impose une autorisation écrite au-delà de 20 millions de dollars annuels, assortie de dix-neuf catégories d'usages prohibés. Par ailleurs, le modèle haut de gamme d'Alibaba, Qwen 3.8-Max, ne dispose d'aucun dépôt public : la stratégie chinoise consiste à ouvrir largement le milieu de gamme tout en réservant le haut de gamme.

L'ouverture n'est donc pas un attribut binaire mais un continuum, et ce continuum se déplace — dans les deux camps — vers la fermeture.

V.5 Le rachat de Hugging Face par Nvidia

Le 26 août 2026, The Information a révélé les discussions avancées entre Nvidia et Hugging Face ; l'accord définitif a été annoncé le 3 septembre 2026 pour 12,93 milliards de dollars. Hugging Face, cofondée par des entrepreneurs français, avait été valorisée 4,5 milliards de dollars en 2023 et réalisait environ 150 millions de dollars de chiffre d'affaires annuel ; elle avait refusé fin 2025 un investissement de 500 millions de Nvidia sur une valorisation de 7 milliards.

Jensen Huang a déclaré que la plateforme « resterait ouverte à l'ensemble de l'écosystème de l'IA ». L'opération doit passer un examen antitrust. Les commentaires relevés soulignent que l'annonce ne précise ni qui détiendra les droits de décision sur les contenus hébergés, ni ce qui se produirait si la politique américaine exigeait une modification de la disponibilité des modèles chinois.

Cette opération mérite d'être traitée comme un fait de souveraineté à part entière. L'infrastructure de distribution des modèles ouverts — l'endroit où l'Europe s'approvisionne — passe sous contrôle du fournisseur dominant des accélérateurs. Si Hugging Face est le supermarché des poids ouverts, il vient d'être racheté par le propriétaire de la centrale électrique.

VI. Cartographie des acteurs et du débat

VI.1 Les acteurs

Les fournisseurs de couche matérielle. Nvidia (accélérateurs, désormais plateforme de distribution), AMD, Samsung et SK Hynix (mémoire), TSMC (fonderie), ASML (lithographie), Huawei et SMIC (filiale chinoise), SiPearl et VSORA (tentatives européennes).

Les laboratoires de frontière fermés. OpenAI, Anthropic, Google DeepMind, désormais Meta Superintelligence Labs. Les valorisations citées en août 2026 — de l'ordre de 965 milliards de dollars pour Anthropic et 852 milliards pour OpenAI — sont des constructions de tours de financement privés et doivent être traitées comme telles.

Les laboratoires à poids ouverts. Alibaba (Qwen), DeepSeek, Moonshot AI (Kimi), Z.ai (GLM), MiniMax côté chinois ; Mistral AI côté européen ; Google (Gemma) et, de manière épisodique, OpenAI (gpt-oss) côté américain.

Les États et institutions. Commission européenne (norme, InvestAI, CADA), EuroHPC (infrastructure), États-Unis (contrôle des exportations, plan d'action), Chine (politique industrielle, étiquetage, enregistrement des modèles), Corée du Sud, Japon, Inde, Émirats arabes unis et Arabie saoudite (capital et infrastructure).

Les acheteurs publics. En France : DINUM, ministère des Armées (accord-cadre notifié à Mistral AI le 16 décembre 2025, piloté par l'AMIAD), Éducation nationale, collectivités territoriales, hôpitaux.

Les régulateurs et évaluateurs. CNIL (future autorité de surveillance du marché au titre du règlement sur l'IA), ANSSI (SecNumCloud), INESIA, Cour des comptes, inspections générales, CEPD et CEPD-EDPS au niveau européen.

VI.2 Quatre positions dans le débat

La position normative. La souveraineté passe par le droit : le règlement sur l'IA, le RGPD et le futur règlement cloud et IA constituent un avantage comparatif européen, car quiconque veut vendre en Europe doit s'y conformer. Objection : l'effet Bruxelles suppose un marché suffisamment attractif pour que l'on accepte d'en payer le prix de conformité, et il n'a jamais produit de champion.

La position industrialiste. La souveraineté passe par le capital et l'infrastructure : financer des champions, construire des gigafactories, réserver la commande publique. C'est la position de la commission d'enquête de l'Assemblée nationale, et celle du rapport « Buy European Tech Act » de la chaire numérique de Sciences Po. Objection : le coût est considérable et la fenêtre étroite.

La position dérégulatrice. La souveraineté passe par la vitesse : c'est la position exprimée par Arthur Mensch devant l'Assemblée nationale — « Réguler pour se défendre, en général, ça ne marche pas. Cela favorise les plus gros acteurs » — et celle qui a inspiré l'omnibus numérique, adopté par le Parlement européen le 16 juin 2026. Objection, formulée par le CEPD et le CEPD-EDPS dans leur avis conjoint du 11 février 2026 : la simplification ne doit pas se payer en droits fondamentaux.

La position des communs. La souveraineté passe par l'ouverture : des poids publiés, des données documentées, des évaluations reproductibles. C'est la position de la fondation Current AI, dotée de 400 millions et lancée au sommet de Paris en février 2025, et celle qu'a portée la déclaration de New Delhi en février 2026. Objection : l'ouverture ne suffit pas à produire l'indépendance, comme le montre le fait que les poids les plus utilisés soient désormais chinois.

Ces quatre positions ne s'excluent pas. Elles se hiérarchisent différemment selon l'horizon considéré : la norme protège à court terme, l'ouverture permet un rattrapage à moyen terme, l'infrastructure seule produit une capacité à long terme.

VII. Une distinction structurante : conformité et capacité

Nexus propose de séparer deux objets que le débat public confond systématiquement.

La **souveraineté de conformité** est la capacité à garantir, par le droit et le contrat, que les données traitées ne sont pas accessibles à une autorité étrangère et que l'utilisateur conserve la maîtrise de sa configuration. Elle se mesure en clauses, en certifications, en localisations. Elle est atteignable rapidement et à coût modéré. Elle protège contre l'accès non autorisé et contre la captation de données. Elle ne protège pas contre l'interruption.

La **souveraineté de capacité** est la possibilité de concevoir, entraîner et faire fonctionner un système sans l'autorisation d'un tiers. Elle se mesure en accélérateurs disponibles, en gigawatts raccordés, en équipes capables de reproduire un entraînement, en alternatives techniques testées. Elle exige des décennies et des dizaines de milliards d'euros. Elle seule protège contre l'interruption.

Le règlement sur le cloud et l'IA, proposé le 3 juin 2026, formalise cette gradation en quatre niveaux d'assurance : niveau 1, données traitées et stockées dans l'Union ; niveau 2, indépendance du fournisseur à l'égard des pays tiers et absence d'interrupteur à distance ; niveau 3, fournisseur non contrôlé par un pays tiers ; niveau 4, composants non contrôlés par un pays tiers et certification de cybersécurité du plus haut niveau. Selon les estimations avancées dans l'analyse publiée par Lawfare, environ 70 % des marchés relèveraient du niveau 1 — accessible aux filiales européennes des fournisseurs américains —, environ 20 % du niveau 2, moins de 10 % du niveau 3 et environ 1 % du niveau 4.

Autrement dit : le dispositif européen le plus ambitieux jamais proposé en matière de souveraineté numérique réserverait à des fournisseurs non contrôlés par un pays tiers moins d'un dixième des marchés publics concernés. Ce n'est pas une critique du texte, qui arbitre entre sécurité, coût et ouverture commerciale. C'est un rappel de ce que « souveraineté » désigne en pratique dans la politique européenne de 2026 : très majoritairement, de la conformité.

La même grille éclaire l'offre de Mistral AI. Depuis le 11 août 2026, l'entreprise commercialise des points d'accès régionaux — Union européenne et États-Unis — avec une majoration tarifaire de 10 %, et un niveau prioritaire assorti d'un engagement de disponibilité de 99,5 % pour une majoration de 75 %. La souveraineté y est un supplément de prix sur une prestation par ailleurs identique. C'est de la conformité, et elle est utile ; ce n'est pas de la capacité.

Le cas le plus révélateur est l'hébergement, annoncé les 11 et 14 août 2026, du modèle chinois GLM-5.2 de Z.ai sur l'infrastructure de Mistral AI, sans modification, avec les mêmes contrôles régionaux et les mêmes engagements de service que les modèles maison. Le directeur technique de Mistral AI l'a justifié en une phrase : le modèle est à poids ouverts, il n'y avait donc « aucune bonne raison de ne pas le faire ». Du point de vue de la conformité, la démonstration est imparable : le modèle

tourne en Europe, les données ne sortent pas, le client contrôle. Du point de vue de la capacité, la conclusion est inverse : le champion européen distribue l'intelligence d'un laboratoire chinois parce qu'elle est meilleure ou moins chère sur certains usages.

VIII. Le cadre européen

VIII.1 Le règlement sur l'IA et son calendrier

Le règlement (UE) 2024/1689 s'applique par étapes. Les interdictions de pratiques et l'obligation de littératie en IA s'appliquent depuis le 2 février 2025 ; les obligations des fournisseurs de modèles à usage général depuis le 2 août 2025. Depuis le 2 août 2026 s'appliquent l'article 50 relatif à la transparence — information de l'utilisateur qu'il interagit avec un système d'IA, marquage des contenus de synthèse —, la compétence des autorités nationales, le régime de sanctions (jusqu'à 15 millions d'euros ou 3 % du chiffre d'affaires mondial annuel pour les manquements à l'article 50) et les bacs à sable réglementaires.

L'omnibus numérique, adopté par le Parlement européen le 16 juin 2026 par 423 voix contre 57 et 174 abstentions, puis par le Conseil le 29 juin 2026, a reporté les obligations relatives aux systèmes à haut risque : annexe III du 2 août 2026 au 2 décembre 2027, annexe I du 2 août 2027 au 2 août 2028. Le marquage des contenus déjà commercialisés et l'interdiction des systèmes de contenus intimes non consentis s'appliquent au 2 décembre 2026.

En France, la CNIL a inscrit à son programme de travail 2026, publié le 7 avril 2026, la préparation de son rôle d'autorité de surveillance du marché, la clarification des conséquences du RGPD pour les modèles d'IA et la répartition des responsabilités entre concepteurs et déployeurs.

VIII.2 L'exemption open source et ses limites

L'article 53, paragraphe 2, du règlement dispense les fournisseurs de modèles à usage général publiés sous licence libre et ouverte des obligations de documentation technique (annexe XI) et d'information des fournisseurs en aval (annexe XII), sous quatre conditions cumulatives : licence libre et ouverte ; licence autorisant l'accès, l'usage, la modification et la distribution ; poids publiquement disponibles ; architecture et informations d'usage publiquement disponibles.

Trois limites sont essentielles. D'abord, l'exemption ne s'applique pas aux modèles présentant un risque systémique — seuil indicatif de 10^{25} opérations en virgule flottante à l'entraînement — qui restent soumis à l'intégralité de l'article 53 et aux obligations de l'article 55 (tests

antagonistes, évaluation des risques, notification d'incidents, cybersécurité). Ensuite, les obligations relatives au droit d'auteur et au résumé public des données d'entraînement s'appliquent en toute hypothèse. Enfin, les licences les plus diffusées aujourd'hui ne satisfont pas nécessairement les quatre conditions : une licence « modified-MIT » assortie de seuils commerciaux, ou une licence propriétaire assortie d'une revue de sécurité obligatoire, n'est pas une licence libre et ouverte au sens du règlement.

Il en résulte une conséquence peu commentée : à mesure que les licences des modèles les plus utilisés se durcissent, l'exemption open source du règlement européen couvre une part décroissante du marché réel.

VIII.3 InvestAI, usines et gigafactories

Le plan « AI Continent » du 9 avril 2025 mobilise 200 milliards d'euros au titre d'InvestAI, dont 20 milliards pour les gigafactories, prévoit 19 usines d'IA et jusqu'à cinq gigafactories, et vise un triplement de la capacité de centres de données de l'Union en cinq à sept ans. Les stratégies « Apply AI » et « AI in Science », présentées le 8 octobre 2025, mobilisent environ 1 milliard d'euros pour l'adoption sectorielle et créent l'institut virtuel RAISE, avec 600 millions d'euros d'Horizon Europe pour l'accès au calcul. Le plan relevait que 13,5 % des entreprises européennes utilisaient alors l'IA.

L'appel EuroHPC pour les gigafactories, lancé le 30 juillet 2026, prévoit jusqu'à sept installations, un dépôt au 12 novembre 2026, une sélection début 2027 et une mise en service dans les dix-huit mois. En France, le consortium AION, mené par Scaleway, porte une candidature de 288 000 équivalents H100 pour 200 mégawatts.

VIII.4 Le règlement cloud et IA et la commande publique

Outre les quatre niveaux d'assurance décrits plus haut, le débat sur le règlement cloud et IA porte sur la préférence européenne. OVHcloud estime avoir besoin que 15 % de la commande publique soit réservée pour atteindre une échelle compétitive. Des associations professionnelles américaines, canadiennes, japonaises et australiennes ont demandé une révision du texte au nom de la non-discrimination.

La commande publique évolue déjà. Le 20 avril 2026, la Commission européenne a attribué un marché cloud de 180 millions d'euros sur six ans à OVHcloud (avec Clever Cloud et DEEP), Scaleway, S3NS et StackIT. En France, selon la réponse du Gouvernement à une question écrite du Sénat publiée le 28 mai 2026, la commande publique en matière de cloud a représenté 84 millions d'euros en 2025, dont 69 % vers des opérateurs européens et 22 millions d'euros d'offres qualifiées

SecNumCloud. Le 10 février 2026, les ministères de la Santé, de la Réforme de l'État et du Numérique ont annoncé que Microsoft n'hébergerait plus le Health Data Hub, un appel d'offres devant désigner un prestataire qualifié SecNumCloud.

Ces montants doivent être rapportés à l'échelle du problème : 84 millions d'euros de commande publique cloud en France, face à des dépenses cloud européennes estimées à 264 milliards d'euros par an dont 83 % captées par des acteurs américains.

IX. Les positions institutionnelles françaises

La Cour des comptes, dans un rapport publié le 19 novembre 2025, juge « très modeste » la transformation des services publics en matière d'IA. Les feuilles de route ministérielles lancées en février 2025 n'avaient pas produit de résultats substantiels à leur échéance de juin. La DINUM estime à environ 100 millions d'euros par an, à compter de 2026, le coût du déploiement d'infrastructures et de services d'IA à l'échelle de l'État — financement non alloué. La Cour recommande la création d'un secrétariat général à l'IA rattaché au Premier ministre, l'utilisation de la commande publique au profit des solutions françaises et l'appui sur les solutions open source.

La mission conjointe IGF-IGA-IGAS, dans son rapport d'avril 2026 sur le déploiement de l'IA dans les administrations publiques, établit un état des lieux comparé avec treize administrations étrangères. Environ 40 % des agents publics utiliseraient l'IA de manière informelle ; l'inventaire recense 141 produits d'IA dans 80 entités publiques, avec un déploiement très inégal (de 0 à 74 projets selon le ministère) ; 90 % des métropoles sont engagées contre 73 % des communes de moins de 3 500 habitants sans aucun projet. La mission privilégie explicitement « la maîtrise des infrastructures et des données plutôt que l'autarcie technologique » et recommande des modèles frugaux, des clauses contractuelles types garantissant la réversibilité et l'autonomie de l'acheteur, et une trajectoire pluriannuelle d'infrastructures mutualisées. Elle relève que la qualification SecNumCloud représente un surcoût de 20 à 40 %.

Le 16 juin 2026, la DINUM a présenté le plan « Notre IA » : quatre outils souverains — L'Assistant, développé avec Mistral AI, Transcript, DiplolA et Visio — hébergés sur l'infrastructure Outscale qualifiée SecNumCloud. L'Assistant a été testé dix mois auprès de 10 000 agents avant généralisation. Les objectifs affichés sont l'élimination de l'« IA clandestine », des gains allant jusqu'à 16 % sur la synthèse documentaire et une relocalisation complète des compétences numériques d'ici 2027.

Le ministère des Armées a notifié le 16 décembre 2025 à Mistral AI un accord-cadre bénéficiant à l'ensemble des armées, directions et services, ainsi qu'au CEA, à l'ONERA et au SHOM, sous pilotage de l'Agence ministérielle pour l'IA de défense créée le 1er mai 2024.

X. Les autres puissances

États-Unis. Le plan « Winning the Race » de juillet 2025 combine soutien à l'open source, exportation de paquets complets de la pile technologique américaine et exigence que les infrastructures soient construites sur des produits américains exempts de technologies d'adversaires. Les contrôles à l'exportation sont devenus un instrument de négociation commerciale plutôt qu'un dispositif stable.

Chine. Politique industrielle de calcul domestique, encouragement académique à l'open source, étiquetage obligatoire des contenus générés et enregistrement préalable des modèles génératifs. Ouverture large du milieu de gamme, rétention du haut de gamme.

Corée du Sud. Plan industriel décennal annoncé le 29 juin 2026, comprenant un investissement conjoint de Samsung et SK Hynix d'environ 518 milliards de dollars dans des capacités de production de semi-conducteurs avancés, avec l'objectif de contrôler la chaîne complète, des puces mémoire aux usines automatisées. L'entrée de Samsung au capital de Mistral AI s'inscrit dans cette stratégie.

Inde. Le sommet AI Impact de New Delhi, du 17 au 23 février 2026, a réuni 88 pays et organisations autour de la déclaration de New Delhi — signée cette fois par les États-Unis et la Chine, contrairement à la déclaration de Paris de février 2025 signée par 58 pays sans les États-Unis ni le Royaume-Uni. Environ 200 milliards de dollars d'engagements y ont été annoncés. L'Inde a dévoilé le 19 février 2026 trois modèles souverains : deux modèles Sarvam de 30 et 105 milliards de paramètres, Param2 de BharatGen (17 milliards de paramètres) et Vachana TTS de Gnani.ai couvrant douze langues indiennes.

Golfe. Les Émirats arabes unis et l'Arabie saoudite sont devenus des pourvoyeurs de capital et d'infrastructure. MGX est coactionnaire du campus de Fouju. Mistral AI a annoncé le 24 août 2026 un partenariat avec HUMAIN, filiale du fonds souverain saoudien, portant sur plusieurs centaines de millions d'euros et incluant le développement de modèles en langue arabe. Le champion européen de la souveraineté exporte donc son modèle de souveraineté, et se finance en partie par des capitaux souverains extra-européens.

XI. Les neuf thèses falsifiables

THESE-HS-2026-010

La souveraineté en IA se décompose en six couches, et l'Europe n'en maîtrise réellement qu'une (Thèse consolidée)

Énoncé. La souveraineté en intelligence artificielle n'est pas un état binaire mais une position sur six couches : accélérateurs de calcul, centres de données et énergie, cloud, modèles et licences, données, normes et talents. L'Union européenne occupe une position dominante sur la seule couche des normes, une position de force sur un maillon amont (la lithographie avec ASML), et une position marginale sur le calcul : les États-Unis concentraient en mai 2025 environ 75 % de la performance mondiale des grappes GPU, la Chine environ 15 %, le reste du monde environ 10 %.

Appuis empiriques. Données Epoch AI sur la répartition de la performance des grappes (mai 2025) ; AI Index 2026 de Stanford HAI (17,1 millions d'équivalents H100 mondiaux, 29,6 GW de capacité électrique dédiée) ; absence d'accélérateur d'entraînement européen en production, le processeur SiPearl Rhea1 étant un processeur généraliste ; part de plus de 70 % des trois hyperscalers américains sur le marché cloud européen.

Robustesse empirique. Élevée sur le calcul et le cloud (sources convergentes, ordres de grandeur stables) ; moyenne sur les talents (indicateurs contradictoires).

Conditions de réfutation. Réfutée si des mesures indépendantes (Epoch AI, Stanford HAI, Commission européenne) établissent une part européenne supérieure à 20 % de la performance mondiale des grappes de calcul, ou si un accélérateur d'entraînement conçu en Europe atteint 10 % du parc installé européen d'ici 2030. Fragilisée si la décomposition en six couches s'avère non pertinente, par exemple si une couche devient substituable aux autres.

Statut. Thèse consolidée

La levée de septembre 2026 consolide un champion industriel sans réduire la dépendance matérielle (Thèse consolidée)

Énoncé. La série D de 3 milliards d'euros annoncée par Mistral AI le 8 septembre 2026, à une valorisation post-monnaire supérieure à 21 milliards d'euros, consolide un acteur industriel de niveau mondial mais ne modifie pas la dépendance de la France et de l'Europe aux accélérateurs américains. Les capacités financées reposent sur du matériel Nvidia.

Appuis empiriques. Annonce du 8 septembre 2026 ; plus de 7,5 milliards d'euros levés au total ; plus de 125 clients grands comptes dans vingt pays ; revenu récurrent annualisé supérieur à 400 millions de dollars (Davos, 12 février 2026), multiplié par vingt en un an ; centre de données de Bruyères-le-Châtel de 44 MW équipé d'environ 13 800 accélérateurs Nvidia GB300 ; campus de Fouju de 1,4 GW développé avec Nvidia et MGX ; objectif de 1 GW de calcul en Europe d'ici 2030 ; accord pluriannuel d'infrastructure GPU avec Microsoft.

Robustesse empirique. Élevée. Les montants et les configurations matérielles sont documentés par les communiqués des parties.

Conditions de réfutation. Réfutée si Mistral AI entraîne un modèle de frontière sur des accélérateurs non américains, ou si la part des accélérateurs américains dans les capacités qu'elle exploite en Europe descend sous 50 % d'ici 2030. Fragilisée si un fournisseur européen ou non américain devient une alternative technique crédible pour l'entraînement.

Statut. Thèse consolidée

Le capital du champion européen est devenu majoritairement extra-français, sans instrument de contrôle correspondant (Hypothèse de travail)

Énoncé. Le contrôle capitalistique de Mistral AI est fragmenté entre des investisseurs dont aucun n'est français, avec trois fournisseurs stratégiques au capital (ASML pour la lithographie, Nvidia pour les accélérateurs, Samsung pour la mémoire). L'État français ne dispose d'aucun instrument juridique spécifique permettant de s'opposer à une évolution de l'actionnariat ou à une cession d'actifs stratégiques, et la recommandation parlementaire d'instituer une action spécifique n'a pas été suivie d'effet.

Appuis empiriques. ASML détient environ 11 % du capital depuis septembre 2025 et dispose d'un siège au comité stratégique ; Samsung mène le tour de septembre 2026 pour un montant estimé par la presse à environ 1 milliard d'euros ; les cofondateurs détiendraient ensemble environ 35,5 % du capital et les investisseurs européens « un peu plus de 62 % » avant le dernier tour (estimations journalistiques) ; recommandation d'une action spécifique de 5 % par la commission d'enquête de l'Assemblée nationale (rapport adopté le 8 juillet 2026, publié le 15 juillet 2026), fondée sur l'article 31-1 de l'ordonnance du 20 août 2014, non mise en œuvre.

Robustesse empirique. Moyenne. Les participations individuelles ne sont pas publiques ; les pourcentages reposent sur des enquêtes de presse non confirmées par l'entreprise.

Conditions de réfutation. Réfutée si l'État institue une action spécifique ou un dispositif équivalent, ou si l'entreprise publie une répartition de son capital et de ses droits de vote démontrant un contrôle français ou européen consolidé. Fragilisée si les mécanismes de contrôle des investissements étrangers en France s'avèrent suffisants en pratique, ou si les statuts comportent des protections non publiques.

Statut. Hypothèse de travail

THESE-HS-2026-013

L'écosystème des modèles à poids ouverts a basculé vers la Chine (Thèse consolidée)

Énoncé. Depuis 2025, les laboratoires chinois dominent la publication et la diffusion des modèles à poids ouverts, tant en volume de téléchargements qu'en taille des modèles publiés et en usage mesuré sur les plateformes d'agrégation.

Appuis empiriques. Rapport « State of Open Models: Summer 2026 » de Hugging Face (14 août 2026) : Qwen au-delà de 3 milliards de téléchargements sur six mois, contre 418 millions pour Google et 227 millions pour Meta ; plafond mensuel de taille des modèles chinois entre 754 et 2 780 milliards de paramètres contre moins de 130 milliards pour les modèles américains cinq mois sur sept ; OpenRouter, 34 250 milliards de jetons consommés par des modèles chinois sur la semaine du 3 au 9 août 2026, quinzième semaine consécutive de dépassement ; Kimi K3 (2 800 milliards de paramètres, 16 juillet 2026) ; DeepSeek V4 sous licence MIT (24 avril 2026) ; dérivés de Qwen représentant plus de 40 % des nouvelles variantes publiées.

Robustesse empirique. Élevée sur la direction du basculement (sources indépendantes convergentes) ; faible sur les niveaux précis (les parts d'usage varient de 30 % à plus de 60 % selon les périmètres et les

périodes).

Conditions de réfutation. Réfutée si un laboratoire non chinois reprend et conserve, pendant deux trimestres consécutifs, la première place en téléchargements et en jetons consommés sur les principales plateformes. Fragilisée si les mesures de téléchargement s'avèrent non représentatives des déploiements en production, ce que des données d'entreprise viendraient établir.

Statut. Thèse consolidée

THESE-HS-2026-014

Le repli américain de l'ouverture est un choix économique, non une contrainte technique ou sécuritaire établie (Hypothèse de travail)

Énoncé. Le retrait de Meta et le caractère ponctuel de l'ouverture d'OpenAI s'expliquent d'abord par la nécessité de monétiser des investissements massifs, et non par une impossibilité technique ni par une démonstration que l'ouverture des poids créerait un risque net supérieur à ses bénéfices.

Appuis empiriques. Publication par Meta de Muse Spark en avril 2026, modèle propriétaire sans poids téléchargeables, alors que la politique publique américaine (« America's AI Action Plan », 23 juillet 2025) encourage explicitement les modèles ouverts ; licences Llama 4 non approuvées par l'Open Source Initiative, avec seuil de 700 millions d'utilisateurs actifs mensuels ; absence de successeur de base à gpt-oss (publié le 5 août 2025) à la mi-2026 ; refus de Meta de signer le code de bonnes pratiques européen (18 juillet 2025), signé par OpenAI et Anthropic.

Robustesse empirique. Moyenne. Les motivations internes ne sont pas directement documentées ; l'argument repose sur la chronologie et sur la contradiction entre politique publique et comportement d'entreprise.

Conditions de réfutation. Réfutée si Meta ou OpenAI publient d'ici 2028 un modèle de frontière sous licence approuvée par l'Open Source Initiative, ce qui invaliderait l'hypothèse d'une fermeture structurelle. Réfutée également si des évaluations indépendantes établissent que l'ouverture des poids de modèles de frontière crée un risque net supérieur à ses bénéfices, ce qui validerait la justification sécuritaire plutôt qu'économique.

Statut. Hypothèse de travail

L'ouverture chinoise se referme elle aussi, par le haut de gamme et par les licences (Hypothèse de travail)

Énoncé. Le modèle chinois d'ouverture n'est ni absolu ni stable : les laboratoires chinois réservent leur haut de gamme, introduisent des seuils commerciaux dans leurs licences et commencent à retenir des poids pour motif de sécurité. L'open weights chinois est un continuum en mouvement vers la fermeture, non un acquis.

Appuis empiriques. Report de deux semaines de la publication des poids de GLM-5.3 annoncé le 18 août 2026 pour renforcement des contrôles de sécurité, le modèle atteignant 84,5 % sur CyberGym et ayant identifié 2 436 vulnérabilités sur 269 projets ; publication le 25 août 2026 sous licence propriétaire « glm-5.3 » avec revue de sécurité obligatoire au-delà de 10 milliards de dollars de chiffre d'affaires ; licences « modified-MIT » de Kimi K2.5 et K3 avec seuils de 100 millions d'utilisateurs actifs mensuels et 20 millions de dollars ; licence MiniMax M3 avec autorisation écrite au-delà de 20 millions de dollars et dix-neuf catégories d'usages prohibés ; absence de dépôt public pour Qwen 3.8-Max.

Robustesse empirique. Moyenne. Les signaux sont récents (été 2026) et peu nombreux ; ils peuvent relever d'ajustements ponctuels plutôt que d'une tendance.

Conditions de réfutation. Réfutée si, sur les exercices 2027 et 2028, la part des modèles chinois de plus de 20 milliards de paramètres publiés sous licence Apache 2.0 ou MIT se maintient au-dessus de 80 % et qu'aucun modèle phare n'est retenu ou soumis à autorisation. Fragilisée si les seuils commerciaux observés s'avèrent sans effet pratique sur les utilisateurs européens.

Statut. Hypothèse de travail

L'open weights est le seul levier de rattrapage européen à court terme, mais il déplace la dépendance au lieu de la supprimer (Hypothèse de travail)

Énoncé. Pour l'Europe, l'adoption de modèles à poids ouverts permet de relocaliser l'inférence, d'auditer partiellement les systèmes et d'éviter l'enfermement propriétaire, à un coût très inférieur à celui d'un entraînement de frontière. Mais elle substitue une dépendance à une

autre : les poids les plus performants sont produits ailleurs, les données d'entraînement restent inconnues et la feuille de route du modèle échappe à l'utilisateur.

Appuis empiriques. Hébergement du modèle chinois GLM-5.2 sur l'infrastructure de Mistral AI annoncé les 11 et 14 août 2026, justifié par le fait qu'il est à poids ouverts ; points d'accès régionaux avec majoration de 10 % ; analyse de l'Institut Montaigne du 7 septembre 2026 identifiant des risques de vulnérabilités intentionnelles et d'opacité des données d'entraînement ; obligation de résumé public des données d'entraînement maintenue par le règlement européen même pour les modèles ouverts ; classement de la commission américaine US-China Economic and Security Review (mars 2026) qualifiant la position chinoise d'avantage auto-renforçant.

Robustesse empirique. Moyenne. La partie « levier de rattrapage » est bien documentée ; la partie « déplacement de la dépendance » repose sur une analyse de risque et non sur des incidents avérés.

Conditions de réfutation. Réfutée si des audits indépendants établissent qu'un modèle à poids ouverts peut être pleinement caractérisé sans accès aux données d'entraînement, ce qui supprimerait l'asymétrie informationnelle. Réfutée également si l'Union finance et livre un modèle de frontière ouvert au niveau des meilleurs modèles mondiaux, ce qui rendrait la dépendance réversible. Fragilisée si des incidents de sécurité imputables à des modèles ouverts chinois sont documentés à grande échelle, ce qui déplacerait le débat de la dépendance vers le risque.

Statut. Hypothèse de travail

THESE-HS-2026-017

La souveraineté européenne a été redéfinie en conformité juridique et en localisation des données (Thèse consolidée)

Énoncé. Les dispositifs européens et français adoptés ou proposés en 2026 organisent une souveraineté de conformité — résidence des données, clauses contractuelles, certification — et non une souveraineté de capacité. Cette redéfinition est explicite dans l'architecture du règlement cloud et IA comme dans les offres commerciales.

Appuis empiriques. Proposition de règlement cloud et IA du 3 juin 2026 avec quatre niveaux d'assurance, dont environ 70 % des marchés relèveraient du niveau 1 accessible aux filiales européennes de fournisseurs américains et moins de 10 % du niveau 3 ; points d'accès régionaux de Mistral AI facturés avec une majoration de 10 % ; qualification SecNumCloud représentant un surcoût de 20 à 40 % selon

la mission IGF-IGA-IGAS ; commande publique cloud française de 84 millions d'euros en 2025 dont 69 % vers des opérateurs européens, face à des dépenses cloud européennes estimées à 264 milliards d'euros par an ; sortie du Health Data Hub de Microsoft annoncée le 10 février 2026 ; report des obligations haut risque du règlement sur l'IA à décembre 2027 et août 2028.

Robustesse empirique. Élevée sur les dispositifs (textes officiels) ; moyenne sur les parts de marché par niveau d'assurance (estimations d'analystes).

Conditions de réfutation. Réfutée si les niveaux 3 et 4 deviennent majoritaires dans les marchés publics européens portant sur des données sensibles, ou si la part des fournisseurs européens dépasse 30 % du marché cloud de l'Union. Fragilisée si la souveraineté de conformité démontre, lors d'un incident réel d'interruption ou d'accès, une efficacité équivalente à celle d'une souveraineté de capacité.

Statut. Thèse consolidée

THESE-HS-2026-018

Le maillon faible français n'est pas le modèle mais la diffusion dans les territoires (Hypothèse de travail)

Énoncé. La France dispose d'un champion des modèles, d'un atout énergétique et d'une offre publique souveraine, mais l'adoption reste concentrée sur les grandes organisations et les grandes villes. L'écart d'usage entre grandes et petites structures, et entre métropoles et communes rurales, constitue le principal risque de souveraineté à dix ans, car une capacité inemployée n'est pas une capacité.

Appuis empiriques. Insee (21 juillet 2026) : 18 % des entreprises implantées en France utilisent l'IA en 2025 contre 10 % en 2024, avec 15 % pour les entreprises de 10 à 49 salariés et 58 % au-delà de 250 salariés, pour une moyenne de l'Union à 20 % ; Bpifrance (mars 2026) : 47 % des PME ont lancé au moins un projet contre 95 % des entreprises de plus de 200 salariés, principal frein le recrutement (60 %) ; mission IGF-IGA-IGAS (avril 2026) : 90 % des métropoles engagées contre 73 % des communes de moins de 3 500 habitants sans aucun projet, environ 40 % d'agents en usage informel, 141 produits d'IA dans 80 entités ; Observatoire Data Publica (14 novembre 2025) : 84 % des collectivités de plus de 3 500 habitants utilisent l'IA générative contre 52 % en 2024, 82 % placent la localisation des données parmi leurs critères, freins principaux le temps et les compétences (62 %) ; Éducation nationale :

outil souverain pour les enseignants financé à hauteur de 20 millions d'euros et attendu pour 2026-2027, adoption enseignante alors inférieure à 20 %.

Robustesse empirique. Élevée sur les écarts d'adoption (sources officielles convergentes) ; moyenne sur le lien causal entre écart d'adoption et perte de souveraineté.

Conditions de réfutation. Réfutée si l'écart d'adoption entre communes de moins de 3 500 habitants et métropoles, et entre entreprises de moins de 50 salariés et entreprises de plus de 250 salariés, se réduit de moitié d'ici 2030 sans politique dédiée — ce qui établirait une diffusion spontanée. Réfutée également si l'adoption massive d'outils non souverains s'avère sans conséquence documentée sur la capacité de décision publique. Fragilisée si les gains de productivité mesurés dans les petites structures s'avèrent négligeables.

Statut. Hypothèse de travail

XII. Implications pour les politiques publiques

Séparer les deux objectifs dans les documents publics. Le premier effet d'une politique de souveraineté est terminologique. Tant que « souveraineté » désigne indistinctement la conformité et la capacité, les arbitrages budgétaires sont impossibles à instruire. Nexus recommande que tout document public distingue explicitement les mesures de conformité (résidence, certification, clauses) et les mesures de capacité (calcul, composants, compétences), avec des indicateurs propres.

Documenter le scénario d'interruption. La question posée par le directeur général de l'ANSSI — que se passe-t-il si l'accès aux technologies américaines est interrompu six mois ? — n'a pas reçu de réponse publique chiffrée. Un exercice de continuité d'activité à l'échelle de l'État, portant sur les fonctions critiques dépendant de fournisseurs extra-européens, constituerait un préalable rationnel à toute décision d'investissement. C'est un exercice de résilience, moins coûteux qu'un programme industriel et immédiatement informatif.

Traiter la file d'attente de raccordement comme un actif public. Les 18 gigawatts de demandes de raccordement enregistrées par RTE en mai 2026 constituent une option stratégique sur la ressource la plus rare en Europe. Les critères d'attribution de ces capacités — ancrage européen, réversibilité, engagement d'usage, contribution à l'équilibre du réseau — relèvent d'une décision politique qui n'a pas fait l'objet d'un débat public à la hauteur de l'enjeu.

Calibrer la préférence européenne sur ce qu'elle peut produire.

OVHcloud estime avoir besoin de 15 % de la commande publique réservée pour atteindre une échelle compétitive. Ce chiffre est vérifiable et discutable ; il a le mérite d'être un ordre de grandeur. Une préférence européenne utile est une préférence dimensionnée sur un seuil de viabilité industrielle, non sur un principe.

Instruire la question de l'action spécifique. La recommandation parlementaire d'une action spécifique de 5 % au capital de Mistral AI mérite une instruction publique — coût, effet dissuasif sur les levées futures, compatibilité avec le droit de l'Union, précédents. Son rejet comme son adoption devraient être motivés. Le silence est le pire des arbitrages.

Financer le calcul académique et le calcul partagé. La mission IGF-IGA-IGAS recommande une trajectoire pluriannuelle d'infrastructures mutualisées. La DINUM chiffre à environ 100 millions d'euros par an le besoin de l'État, non alloué. Ces montants sont modestes au regard des enjeux affichés : l'écart entre le discours et la ligne budgétaire est, en l'état, le principal indicateur de crédibilité de la politique française.

Stabiliser le cadre plutôt que l'alléger. L'omnibus numérique a reporté les obligations haut risque de seize à vingt-quatre mois. Les PME interrogées par Bpifrance citent à 52 % un cadre réglementaire perçu comme flou parmi leurs freins. Un report n'est pas une clarification : la prévisibilité du calendrier compte davantage, pour l'investissement, que son caractère plus ou moins exigeant.

XIII. Implications pour les territoires

L'angle territorial est, pour Nexus, le prolongement nécessaire de l'analyse géopolitique. Une nation ne perd pas sa souveraineté seulement par le haut, en subissant une interruption d'approvisionnement ; elle la perd aussi par le bas, lorsque ses organisations publiques et ses entreprises deviennent incapables d'exercer un choix éclairé.

L'écart est d'abord un écart de taille, pas de territoire. Les données de l'Insee montrent 15 % d'adoption dans les entreprises de 10 à 49 salariés contre 58 % au-delà de 250 salariés. Celles de la mission IGF-IGA-IGAS montrent 90 % de métropoles engagées contre 73 % de communes de moins de 3 500 habitants sans aucun projet. Ces deux écarts se superposent dans les territoires ruraux et les villes moyennes, où la densité de petites structures est la plus forte.

Les collectivités ont déjà intégré le critère de souveraineté. Le baromètre de l'Observatoire Data Publica établit que 82 % des collectivités placent la localisation des données en France ou en Europe parmi leurs critères et que 80 % de celles qui expérimentent ont mis en place des mesures d'IA de confiance. La demande de souveraineté existe donc ; c'est l'offre accessible à cette échelle qui manque.

Le premier chantier est la mutualisation. Une commune de 3 000 habitants ne peut ni négocier une clause de réversibilité, ni auditer un modèle, ni recruter un profil data. Les structures de mutualisation existantes — syndicats mixtes numériques, centres de gestion, agences techniques départementales, intercommunalités — constituent l'échelle pertinente. Les clauses contractuelles types recommandées par la mission IGF-IGA-IGAS auraient un effet démultiplicateur si elles étaient publiées et directement utilisables par ces structures.

Le deuxième chantier est l'« IA clandestine ». L'estimation de 40 % d'agents publics utilisant l'IA de manière informelle décrit un risque de souveraineté concret : des données publiques transitent par des services dont ni le lieu de traitement ni les conditions de réutilisation ne sont maîtrisés. Le plan « Notre IA » traite ce problème à l'échelle de l'État avec des outils hébergés sur une infrastructure qualifiée SecNumCloud ; aucun dispositif équivalent n'existe pour les 35 000 communes et leurs groupements.

Le troisième chantier est la formation. Les freins déclarés sont convergents : 62 % des collectivités et 60 % des PME citent le temps et les compétences. Les IA-Clusters dotés de 360 millions d'euros visent 100 000 personnes formées d'ici 2030 ; ce chiffre est à rapporter aux 5,7 millions d'agents publics et aux millions de salariés de PME. La formation initiale des enseignants et l'outil souverain financé à hauteur de 20 millions d'euros, attendu pour l'année scolaire 2026-2027, constituent le levier de long terme le mieux identifié.

Le quatrième chantier est l'usage de la commande publique comme instrument de diffusion. La commande publique française de cloud représentait 84 millions d'euros en 2025, dont 69 % vers des opérateurs européens. Étendre ce suivi aux achats d'IA, publier la part des solutions européennes et la rendre opposable dans les délibérations locales transformerait un principe en indicateur — et un indicateur en levier.

XIV. Angles morts

Le coût réel de la souveraineté n'est pas chiffré. Aucune évaluation publique française ne chiffre le coût complet d'une capacité d'entraînement souveraine — accélérateurs, énergie, équipes,

renouvellement à trois ans. Sans ce chiffre, le débat oppose des principes à des principes.

L’empreinte environnementale est traitée séparément. Les 18 gigawatts de demandes de raccordement, les 29,6 gigawatts de capacité mondiale recensés par Stanford HAI et la trajectoire de 23 à 28 térawattheures en 2035 pour la France n’ont pas fait l’objet d’une articulation explicite avec la Stratégie nationale bas carbone. Le hors-série Nexus consacré au réchauffement climatique et le présent hors-série se croisent ici sans que la littérature publique ne les articule.

La sécurité des modèles ouverts n’est pas évaluée publiquement. Les risques évoqués — vulnérabilités intentionnelles, comportements déclenchés, exfiltration — sont plausibles et invérifiables par les utilisateurs. L’INESIA, créé en 2025, a vocation à combler ce vide ; aucune évaluation publique de modèles ouverts chinois n’est disponible à ce jour en France.

La dépendance logicielle intermédiaire est ignorée. Le débat porte sur les accélérateurs et les modèles. Il ignore la couche logicielle intermédiaire — CUDA, bibliothèques d’inférence, outils d’orchestration, plateformes de distribution — dont le rachat de Hugging Face par Nvidia rappelle qu’elle est un point de contrôle.

L’effet de l’IA sur l’emploi n’est pas documenté en France. La mission IGF-IGA-IGAS retient que 13 à 20 % des agents publics sont significativement exposés ; l’AI Index 2026 relève une baisse de 20 % du nombre de développeurs logiciels de 22 à 25 ans depuis 2024 aux États-Unis. Aucune étude française équivalente n’est publiée.

Les données d’usage réel manquent. Toutes les mesures de diffusion des modèles reposent sur des plateformes de développeurs. La part des modèles chinois, américains et européens dans les systèmes de production des entreprises et des administrations européennes n’est, à notre connaissance, mesurée par personne.

XV. Trois scénarios à l’horizon 2032

Scénario A — Consolidation européenne. Les gigafactories sélectionnées début 2027 entrent en service en 2028-2029 ; le règlement cloud et IA est adopté avec des seuils de niveau 3 significatifs ; la commande publique atteint un seuil de viabilité pour les fournisseurs européens ; Mistral AI atteint son objectif de 1 gigawatt et consolide sa position sur les modèles ouverts occidentaux ; l’adoption dans les PME et les collectivités rejoint la moyenne européenne. La

dépendance matérielle subsiste mais devient gérable : la France passe d'une position de dépendance à une position d'interdépendance négociée.

Scénario B — Conformité sans capacité. Les dispositifs de conformité se déploient ; les niveaux 1 et 2 dominent la commande publique ; les gigafactories prennent du retard ou changent d'échelle ; le capital des champions se dilue davantage ; l'adoption reste concentrée sur les grandes organisations. L'Europe dispose d'un droit respecté et d'aucune capacité propre. C'est la trajectoire tendancielle, et la plus probable en l'absence de décision.

Scénario C — Rupture d'approvisionnement. Une décision américaine ou un conflit sur les détroits asiatiques interrompt l'approvisionnement en accélérateurs ou en mémoire. Les capacités installées continuent de fonctionner ; le renouvellement s'arrête ; l'écart de performance s'accroît de deux à trois générations en trois ans. Les organisations qui auront gardé une pile ouverte et documentée pourront continuer ; les autres seront arrêtées par une clause contractuelle.

La probabilité relative de ces scénarios n'est pas estimable. Leur utilité est de faire apparaître une asymétrie : les mesures qui préservent la valeur dans le scénario C — poids ouverts, réversibilité, compétences internes, calcul mutualisé — sont aussi celles qui coûtent le moins dans les scénarios A et B. Ce sont des mesures sans regret.

XVI. Bibliographie hiérarchisée

Toutes les URL ont été consultées le 12 septembre 2026.

XVI.1 Sources primaires officielles

Assemblée nationale, commission d'enquête sur les dépendances structurelles et les vulnérabilités systémiques dans le secteur du numérique et les risques pour l'indépendance de la France (créée le 3 février 2026 ; rapport adopté le 8 juillet 2026 ; publié le 15 juillet 2026).
<https://www.assemblee-nationale.fr/dyn/17/organes/cnpe/ce-dependance-numerique>

Commission européenne, « AI Continent Action Plan » (9 avril 2025).
https://commission.europa.eu/topics/competitiveness/ai-continent_en

Commission européenne, stratégies « Apply AI » et « AI in Science » (8 octobre 2025).
https://france.representation.ec.europa.eu/informations/la-commission-lance-deux-strategies-pour-accelerer-ladoption-de-lia-dans-lindustrie-et-la-science-2025-10-08_fr

Commission européenne, proposition de règlement sur le développement du cloud et de l'IA (3 juin 2026). <https://digital-strategy.ec.europa.eu/en/policies/cloud-and-ai-development-act>

EuroHPC, lancement de l'appel « AI Gigafactories » (30 juillet 2026). https://www.eurohpc-ju.europa.eu/eurohpc-joint-undertaking-launches-ai-gigafactories-call-2026-07-30_en

Comité européen de la protection des données et Contrôleur européen de la protection des données, avis conjoint sur l'omnibus numérique (11 février 2026). https://www.edpb.europa.eu/news/news/2026/digital-omnibus-edpb-and-edps-support-simplification-and-competitiveness-while_en

Cour des comptes, observations définitives « France Travail et l'intelligence artificielle » (8 janvier 2026). <https://www.ccomptes.fr/fr/publications/france-travail-et-lintelligence-artificielle>

IGF, IGA et IGAS, « Le déploiement de l'intelligence artificielle dans les administrations publiques » (avril 2026). [https://igas.gouv.fr/sites/igas/files/2026-07/D%C3%A9ploiement%20IA%20\(2026\).pdf](https://igas.gouv.fr/sites/igas/files/2026-07/D%C3%A9ploiement%20IA%20(2026).pdf)

DINUM, plan « Notre IA » pour les services publics (16 juin 2026). <https://www.numerique.gouv.fr/sinformer/espace-presse/intelligence-artificielle-ia-utile-humaine-souveraine-services-publics/>

CNIL, programme de travail 2026 (7 avril 2026). <https://www.cnil.fr/fr/accompagnement-des-professionnels-le-programme-de-travail-de-la-cnil-pour-2026>

Sénat, question écrite de Mme Cathy Apurceau-Poly et réponse du Gouvernement (12 mars et 28 mai 2026). <https://www.senat.fr/questions/base/2026/qSEQ260308036.html>

Ministère de l'Éducation nationale, mesures relatives à l'IA et à l'éducation (7 février 2025). <https://www.education.gouv.fr/intelligence-artificielle-au-service-de-l-education-des-mesures-ambitieuses-pour-accompagner-les-462591>

Gouvernement français, « IA : une nouvelle impulsion pour la stratégie nationale » (6 février 2025). <https://www.info.gouv.fr/actualite/ia-une-nouvelle-impulsion-pour-la-strategie-nationale>

Insee, « Les technologies de l'information et de la communication dans les entreprises en 2025 », Insee Première n° 2120 (21 juillet 2026). <https://www.insee.fr/fr/statistiques/9025878>

RTE, « Les data centers en chiffres clés » (mai 2026). <https://www.rte-france.com/bases-electricite/consommation-electricite/essor-data-centers-france>

Maison-Blanche, « Winning the Race: America's AI Action Plan » (23 juillet 2025). <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>

Ministère indien des Affaires étrangères, adoption de la déclaration de New Delhi (23 février 2026). <https://www.mea.gov.in/press-releases.htm>

XVI.2 Publications d'entreprises et de plateformes

Mistral AI, « Introducing Mistral 3 » (2 décembre 2025). <https://mistral.ai/news/mistral-3/>

Mistral AI, levée de 1,7 milliard d'euros menée par ASML (9 septembre 2025). <https://mistral.ai/fr/news/mistral-ai-raises-1-7-b-to-accelerate-technological-progress-with-ai>

Mistral AI, « AI Now Summit 2026 » (28 mai 2026). <https://mistral.ai/news/ai-now-summit-2026/>

MGX, « Bpifrance, Mistral and MGX to Expand Campus AI across France » (1er juin 2026). <https://www.mgx.ae/news-insights/bpifrance-mistral-and-mgx-expand-campus-ai-across-france-deliver-3-gw-network-ai>

Hugging Face, « State of Open Models: Summer 2026 » (14 août 2026). <https://huggingface.co/blog/state-of-open-models-summer-2026>

OpenAI, « Introducing gpt-oss » (5 août 2025). <https://openai.com/index/introducing-gpt-oss/>

XVI.3 Sources académiques et instituts de recherche

Stanford HAI, « Inside the AI Index: 12 Takeaways from the 2026 Report » (13 avril 2026). <https://hai.stanford.edu/news/inside-the-ai-index-12-takeaways-from-the-2026-report>

Epoch AI, « The US hosts the majority of GPU cluster performance, followed by China » (mai 2025). <https://epoch.ai/data-insights/ai-supercomputers-performance-share-by-country>

Institut Montaigne, François Godement, « Offensive chinoise en Europe : l'open source, nouveau cheval de Troie numérique » (7 septembre 2026). <https://institutmontaigne.org/expressions/offensive-chinoise-en-europe-lopen-source-nouveau-cheval-de-troie-numerique>

Sciences Po, chaire numérique, gouvernance et souveraineté, rapport « Buy European Tech Act » (juin 2024). https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2024/06/Rapport-Buy-European-Tech-Act_compressed.pdf

Lawfare, « The EU Cloud and AI Development Act » (juin 2026). <https://www.lawfaremedia.org/article/the-eu-cloud-and-ai-development-act>

XVI.4 Presse spécialisée

TechCrunch, « Mistral raises €3B as sovereign AI becomes big business » (8 septembre 2026). <https://techcrunch.com/2026/09/08/mistral-raises-e3b-as-sovereign-ai-becomes-big-business/>

TechCrunch, « Nvidia closes in on Hugging Face acquisition » (26 août 2026). <https://techcrunch.com/2026/08/26/nvidia-closes-in-on-hugging-face-acquisition/>

MIT Technology Review, « What's next for Chinese open-source AI » (12 février 2026). <https://www.technologyreview.com/2026/02/12/1132811/whats-next-for-chinese-open-source-ai/>

Le MagIT, « ASML investit 1,3 milliard d'euros dans Mistral AI » (9 septembre 2025). <https://www.lemagit.fr/actualites/366630303/ASML-investit-13-milliard-deuros-dans-Mistral-AI>

Le MagIT, « Processeur : le Rhea1 équipera les supercalculateurs dès fin 2026 » (27 mai 2026). <https://www.lemagit.fr/actualites/366643528/Processeur-le-Rhea1-equipera-les-supercalculateurs-des-fin-2026>

Le MagIT, « Cloud souverain : Bruxelles sélectionne OVHcloud, Scaleway et S3NS » (20 avril 2026). <https://www.lemagit.fr/actualites/366641900/Cloud-souverain-Bruelles-selectionne-OVHcloud-Scaleway-et-S3NS>

Techzine, « Mistral to host external models, Chinese Z.ai to be first » (14 août 2026). <https://www.techzine.eu/news/analytics/143617/mistral-to-host-external-models-chinese-z-ai-to-be-first/>

Digital Applied, « We Read the Licences on 2026 Open-Weight Models » (16 août 2026). <https://www.digitalapplied.com/blog/open-weight-model-licence-audit-2026>

Digital Applied, « Mistral Hosts Rival Open Models on Regional Endpoints » (11 août 2026). <https://www.digitalapplied.com/blog/mistral-hosts-rival-open-models-regional-endpoints-2026>

The New Stack, « Meta abandons open-source Llama for proprietary Muse Spark » (30 avril 2026). <https://thenewstack.io/meta-abandons-llama-spark/>

VentureBeat, « China's Moonshot AI releases Kimi K3 » (16 juillet 2026). <https://venturebeat.com/technology/chinas-moonshot-ai-releases-kimi-k3-the-largest-open-source-model-ever-rivaling-top-u-s-systems>

DataCamp, « DeepSeek V4: Features, Benchmarks, and Comparisons » (avril 2026). <https://www.datacamp.com/blog/deepseek-v4>

Developpez.com, report de la publication des poids de GLM-5.3 (18 août 2026). <https://intelligence-artificielle.developpez.com/actu/386247/>

Datacenter Dynamics, « MGX, Bpifrance, Nvidia, and Mistral AI plan 1.4GW Paris data center campus » (20 mai 2025).

<https://www.datacenterdynamics.com/en/news/mgx-bpifrance-nvidia-and-mistral-ai-plan-14gw-paris-data-center-campus/>

Acteurs publics, « IA : la Cour des comptes estime que les efforts de transformation publique sont encore trop modestes » (19 novembre 2025). <https://acteurspublics.fr/articles/ia-la-cour-des-comptes-estime-que-les-efforts-de-transformation-publique-sont-encore-trop-modestes/>

Introl, « BIS H200 Export Policy Shift & AI OVERWATCH Act » (2026). <https://introl.com/blog/bis-h200-china-export-policy-ai-overwatch-act-2026>

Banque des Territoires, « Observatoire Data Publica : l'IA générative poursuit sa percée » (14 novembre 2025). <https://www.banquedesterritoires.fr/observatoire-data-publica-lia-generative-poursuit-sa-percee-et-impacte-les-collectivites-de-toutes>

IREDIC, « Mistral AI au service de la Défense » (9 février 2026). <https://iredic.fr/2026/02/09/mistral-ai-au-service-de-la-defense-un-accord-cadre-entre-souverainete-technologique-et-dependance-strategique/>

XVI.5 Presse générale

Fortune, « India's AI Impact Summit closes with the New Delhi Declaration and a \$200 billion boost » (23 février 2026). <https://fortune.com/2026/02/23/indias-ai-impact-summit-closes-with-the-new-delhi-declaration-and-a-200-billion-boost/>

South China Morning Post, « Alibaba's Qwen family captures over 50 % of global open-source downloads » (avril 2026). <https://www.scmp.com/tech/big-tech/article/3349552/alibabas-qwen-family-captures-over-50-global-open-source-downloads-report-finds>

LCP, adoption du rapport de la commission d'enquête sur les vulnérabilités numériques (8 juillet 2026). <https://lcp.fr/actualites/numerique-le-rapport-de-la-commission-d-enquete-sur-les-vulnerabilites-de-la-france-a>

La Presse, « Intelligence artificielle : la Corée du Sud annonce un plan colossal » (29 juin 2026). <https://www.lapresse.ca/affaires/techno/2026-06-29/intelligence-artificielle/la-coree-du-sud-annonce-un-plan-colossal.php>

Clubic, « Mistral AI est-elle encore une entreprise française ? » (septembre 2026). <https://www.clubic.com/actualite-628750-mistral-ai-est-elle-encore-une-entreprise-francaise-enquete-sur-qui-sont-ses-investisseurs-d-ici-et-d-ailleurs.html>

XVII. Glossaire

Accélérateur — Circuit spécialisé dans les calculs matriciels requis par l'apprentissage profond. Les accélérateurs graphiques (GPU) de Nvidia dominent le segment de l'entraînement.

AI Act — Nom usuel du règlement (UE) 2024/1689 établissant des règles harmonisées concernant l'intelligence artificielle.

AMIAD — Agence ministérielle pour l'intelligence artificielle de défense, créée le 1er mai 2024.

CADA — Cloud and AI Development Act, proposition de règlement européen présentée le 3 juin 2026, établissant quatre niveaux d'assurance de souveraineté pour les services en nuage.

Cloud Act — Loi américaine de 2018 permettant aux autorités américaines d'exiger d'un fournisseur relevant de leur juridiction la communication de données, y compris stockées hors des États-Unis.

Équivalent H100 — Unité de compte permettant de comparer des parcs d'accélérateurs hétérogènes, rapportée aux performances de l'accélérateur Nvidia H100.

Gigafactory d'IA — Installation de calcul de très grande échelle destinée à l'entraînement de modèles de fondation, dans la terminologie de la Commission européenne.

GPAI — General Purpose AI, modèle d'intelligence artificielle à usage général, catégorie réglementaire du règlement européen.

Inférence — Exécution d'un modèle déjà entraîné pour produire une réponse, par opposition à l'entraînement.

INESIA — Institut national pour l'évaluation et la sécurité de l'intelligence artificielle, annoncé en février 2025.

Modèle de fondation — Modèle entraîné à grande échelle sur des données massives, réutilisable pour de nombreuses tâches en aval.

Open source / open weights — L'open source suppose la mise à disposition du code et des conditions de reproduction ; les open weights désignent la seule publication des paramètres du modèle. La plupart des modèles dits « ouverts » sont des open weights.

OSI — Open Source Initiative, organisation qui approuve les licences répondant à sa définition de l'open source.

Poids — Paramètres numériques d'un modèle, résultat de l'entraînement.

Risque systémique — Qualification appliquée par le règlement européen aux modèles à usage général les plus capables, avec un seuil indicatif de 10^{25} opérations en virgule flottante à l'entraînement.

SecNumCloud — Qualification délivrée par l'ANSSI aux offres de cloud répondant à des exigences de sécurité et de protection contre les législations à portée extraterritoriale.

Souveraineté de capacité — Possibilité de concevoir, entraîner et faire fonctionner un système sans autorisation d'un tiers.

Souveraineté de conformité — Garantie, par le droit et le contrat, que les données traitées ne sont pas accessibles à une autorité étrangère et que l'utilisateur conserve la maîtrise de sa configuration.

Token (jeton) — Unité élémentaire de texte traitée par un modèle de langage ; unité de facturation usuelle des interfaces applicatives.

PARTIE III

Recherche empirique source

RAPPORT DE RECHERCHE EMPIRIQUE

Intelligence artificielle et souveraineté

Collecte empirique — matière première (étape 1)

Domaine : Hors-série

Version 0-1 — Document de travail Nexus

I. Périmètre et enjeux

L'intelligence artificielle est devenue, en moins de quatre ans, un objet de politique étrangère. Ce qui se jouait en 2022 dans des laboratoires de recherche se joue désormais dans des décrets de contrôle des exportations, des accords de fourniture d'accélérateurs graphiques, des raccordements électriques et des tours de table capitalistiques. Le présent document rassemble les faits datés et sourcés mobilisés par le hors-série Nexus « Intelligence artificielle et souveraineté ». Il ne comporte ni thèse ni recommandation : il constitue la base empirique à partir de laquelle le Rapport v2-O formule ses neuf thèses falsifiables.

Le périmètre retenu couvre six couches, qui sont autant de points de dépendance possibles : (1) les semi-conducteurs et les accélérateurs de calcul ; (2) les centres de données et l'énergie qui les alimente ; (3) les couches d'infrastructure logicielle et de cloud ; (4) les modèles eux-mêmes, leurs poids et leurs licences ; (5) les données d'entraînement et d'usage ; (6) les normes, les talents et le droit. La souveraineté n'est pas un état binaire : c'est une position sur chacune de ces six couches, et une capacité à ne pas être arrêté sur aucune d'entre elles.

Trois séquences d'actualité structurent le document. La première est la levée de fonds de Mistral AI de septembre 2026, menée par un industriel coréen, qui pose la question du capital de la souveraineté. La deuxième est le basculement de l'écosystème des modèles à poids ouverts vers la Chine, mesuré par les plateformes de distribution. La troisième est le mouvement inverse observé chez plusieurs acteurs américains, et désormais chez certains laboratoires chinois : une fermeture progressive, par le haut de gamme et par les licences.

II. Méthodologie et hiérarchie des sources

Sources mobilisées, par ordre de priorité. (1) Sources primaires officielles : communiqués et publications de la Commission européenne (plan « AI Continent », stratégie « Apply AI », proposition de règlement CADA), de l'entreprise commune EuroHPC, du Parlement et du Conseil de l'Union européenne (règlement 2024/1689 dit AI Act et son omnibus numérique), de l'Assemblée nationale (commission d'enquête sur les dépendances numériques), du Sénat (questions écrites et réponses du Gouvernement), de la Cour des comptes, des inspections générales IGF-IGA-IGAS, de la DINUM, de la CNIL, de l'ANSSI, du ministère de l'Éducation nationale, de l'Insee, de RTE, du Bureau of Industry and Security américain, du ministère des Affaires étrangères indien, et des

entreprises concernées (Mistral AI, Nvidia, Hugging Face, Z.ai, Moonshot AI, DeepSeek, MGX). (2) Sources académiques et instituts de recherche : Stanford HAI (AI Index 2026), Epoch AI, Institut Montaigne, chaire numérique de Sciences Po. (3) Presse spécialisée : TechCrunch, MIT Technology Review, Le MagIT, L'Usine Digitale, Datacenter Dynamics, VentureBeat, South China Morning Post, Acteurs publics, La Gazette des communes. (4) Presse générale : Bloomberg, CNBC, Fortune, Le Monde Informatique, LCP.

Démarche. (a) Reconstitution chronologique des faits de 2023 à septembre 2026, avec datation systématique ; (b) vérification des faits récents (2025-2026) sur au moins deux sources indépendantes ; (c) distinction explicite entre faits établis (documents officiels, communiqués d'entreprise), estimations (calculs d'analystes, chiffres de marché) et annonces (engagements non encore exécutés) ; (d) construction de tableaux comparatifs par couche de dépendance ; (e) formalisation falsificationniste de neuf thèses dans le Rapport v2-0.

Limites explicitées. (a) La vitesse d'évolution du domaine rend certains chiffres périssables à l'échelle du trimestre : les parts d'usage des modèles, en particulier, varient fortement d'un mois à l'autre. (b) Les montants des levées de fonds privées et les valorisations post-monétaires ne sont pas auditable : ils sont communiqués par les entreprises et leurs investisseurs. (c) Les participations individuelles au capital de Mistral AI ne sont pas publiques ; les pourcentages cités dans la presse sont des estimations. (d) Les métriques de téléchargement et d'usage des modèles ouverts (Hugging Face, OpenRouter) mesurent des comportements de développeurs, non des déploiements en production, et sur-représentent les usages expérimentaux. (e) Les capacités de calcul annoncées (en gigawatts ou en nombre d'accélérateurs) sont des projets, dont l'exécution s'étale sur cinq à dix ans. (f) Aucune source publique fiable ne permet d'établir la part européenne du parc mondial d'accélérateurs installés en 2026 ; les estimations disponibles portent sur les grappes de calcul recensées.

III. La levée de fonds Mistral AI de septembre 2026

III.1 Les faits établis

Le 8 septembre 2026, Mistral AI a annoncé la clôture d'un tour de financement de série D de 3 milliards d'euros, portant sa valorisation post-monétaire au-delà de 21 milliards d'euros (environ 24 milliards de dollars). L'entreprise et la presse économique présentent l'opération comme la plus importante levée en fonds propres jamais réalisée par

une entreprise technologique européenne. Le tour est mené par Samsung Electronics, avec pour co-chefs de file le Scaleup Europe Fund géré par EQT et PSG Equity, déjà investisseur.

Les nouveaux entrants annoncés sont Advent, des fonds gérés par BlackRock et le Grand-Duché de Luxembourg. Les investisseurs historiques qui participent au tour comprennent a16z (Andreessen Horowitz), ASML, Belfius, BNP Paribas CIB, Bpifrance, Carmignac, DST Global, Eurazeo, General Catalyst, Headline, Hillspire, Index Ventures, Korelya Capital, Lightspeed, NVIDIA, le fonds Solar de Phoenix Court et Salesforce Ventures. Le total des fonds propres levés par l'entreprise depuis sa création dépasse 7,5 milliards d'euros.

L'entreprise déclare opérer dans vingt pays et compter plus de 125 clients grands comptes, parmi lesquels Airbus, ASML et HSBC. Elle annonce un objectif de 1 gigawatt de capacité de calcul en Europe d'ici 2030. Selon TechCrunch, l'entrée de Samsung a reçu l'aval des autorités françaises.

III.2 Les estimations et les zones d'incertitude

Le montant investi par Samsung Electronics n'a pas été rendu public. La presse économique française et anglo-saxonne avance un ordre de grandeur d'environ 1 milliard d'euros, soit approximativement un tiers du tour. Ce chiffre doit être traité comme une estimation. Aucune source publique ne donne le pourcentage de capital correspondant, ni l'existence d'un siège au conseil ou au comité stratégique.

La motivation stratégique de Samsung est en revanche documentée : le groupe cherche à élargir sa base de clients pour ses semi-conducteurs de mémoire et à capter la demande liée à la montée en charge des infrastructures d'IA. Samsung est donc à la fois fournisseur potentiel et actionnaire, configuration déjà observée avec ASML en 2025 et avec NVIDIA depuis 2024.

Sur la structure du capital, l'enquête la plus détaillée disponible (Clubic, septembre 2026) indique que les trois cofondateurs détiendraient ensemble environ 35,5 % du capital et que les investisseurs français et européens détiennent, avant le tour de septembre 2026, « un peu plus de 62 % » du capital. Ces chiffres ne sont pas confirmés par l'entreprise et doivent être traités comme des estimations journalistiques.

III.3 Historique des tours de financement

Date	Montant	Valorisation post-monnaire	Chef de file
Juin 2023	105 M€	~240 M€	Lightspeed
Décembre 2023	385 M€	> 2 Md€	Andreessen Horowitz
Juin 2024	600 M€	5,8 Md€	General Catalyst

Date	Montant	Valorisation post-monnaire	Chef de file
9 septembre 2025	1,7 Md€ (série C)	11,7 Md€	ASML (1,3 Md€)
Mars 2026	830 M\$ (dette)	—	Consortium de sept banques
8 septembre 2026	3 Md€ (série D)	> 21 Md€	Samsung Electronics

Le tour de septembre 2025 mérite une attention particulière. ASML, le fabricant néerlandais d'équipements de lithographie, y a investi 1,3 milliard d'euros sur les 1,7 milliard levés, acquérant environ 11 % du capital et devenant ainsi le premier actionnaire individuel de Mistral AI, devant chacun des trois cofondateurs. Roger Dassen, directeur financier d'ASML, siège au comité stratégique. La motivation affichée était l'application de l'IA à l'optimisation des équipements de lithographie extrême ultraviolet.

Le financement de mars 2026, d'un montant de 830 millions de dollars, n'est pas une levée en fonds propres mais un financement par dette apporté par un consortium de sept banques, destiné au centre de données de l'Essonne.

III.4 Chiffre d'affaires et clients

Lors du Forum économique mondial de Davos, le 12 février 2026, Arthur Mensch a annoncé un revenu récurrent annualisé supérieur à 400 millions de dollars, soit une multiplication par vingt en un an (environ 20 millions de dollars un an plus tôt), et une cible d'un milliard de dollars de revenu récurrent à fin 2026. L'effectif était alors estimé entre 350 et 600 personnes selon les sources.

Lors de son audition devant la commission d'enquête de l'Assemblée nationale sur les dépendances numériques, le 12 mai 2026, Arthur Mensch a indiqué que 70 % du chiffre d'affaires de Mistral AI était réalisé hors de France et que la commande publique française représentait environ 10 % de l'activité.

III.5 Infrastructure de calcul

Le premier centre de données détenu par Mistral AI est situé à Bruyères-le-Châtel (Essonne), opéré par Eclairion, d'une capacité de 44 mégawatts, équipé d'environ 13 800 accélérateurs Nvidia GB300, mis en service au deuxième trimestre 2026. Un second site d'inférence de 10 mégawatts, hébergé par Digital Realty aux Ulis, a été annoncé le 28 mai 2026 pour une ouverture au troisième trimestre 2026.

Le projet Campus AI, coentreprise annoncée le 20 mai 2025 entre MGX (fonds émirati), Bpifrance, Mistral AI et NVIDIA, prévoit un campus de 1,4 gigawatt à Fouju (Seine-et-Marne) : début des travaux au second semestre 2026, exploitation à partir de 2028. Le 1er juin 2026, les

partenaires ont annoncé l'extension du projet à un réseau de 3 gigawatts réparti sur plusieurs sites en France, Mistral AI étant locataire d'ancrage avec un accès prioritaire au calcul.

Mistral AI a par ailleurs conclu un accord d'infrastructure GPU pluriannuel de plusieurs milliards de dollars avec Microsoft, et des partenariats avec Vast Data et Dell (systèmes NVIDIA GB300 NVL72).

IV. La bascule de l'open weights vers la Chine

IV.1 Les mesures de distribution

Le rapport « State of Open Models: Summer 2026 » publié par Hugging Face le 14 août 2026 constitue la source la plus complète disponible.

Ses principaux constats :

- La famille Qwen d'Alibaba a franchi les 3 milliards de téléchargements sur six mois, contre 418 millions pour les modèles ouverts de Google et 227 millions pour ceux de Meta sur l'année 2026.
- Le plafond mensuel de taille des modèles ouverts publiés par des laboratoires chinois s'est situé entre 754 milliards et 2 780 milliards de paramètres ; celui des laboratoires américains est resté sous 130 milliards de paramètres cinq mois sur sept.
- La concentration de l'attention est extrême : 85,6 % des dépôts de modèles comptent moins de 200 téléchargements cumulés, et 1,5 % des dépôts concentrent 99,2 % des téléchargements.
- Parmi les modèles chinois de plus de 20 milliards de paramètres, 59 % sont publiés sous licence Apache 2.0 et 22 % sous licence MIT ; les restrictions non commerciales y étaient jusqu'à récemment quasiment absentes.
- La plateforme comptait 2,96 millions de dépôts de modèles et 1 million de jeux de données à l'été 2026.

Une analyse indépendante publiée par la lettre Interconnects AI le 9 avril 2026, reprise par le South China Morning Post, établissait que Qwen représentait plus de 50 % des téléchargements mondiaux de modèles ouverts en mars 2026, avec près d'un milliard de téléchargements cumulés et 153,6 millions pour le seul mois de février 2026.

Sur la plateforme d'agrégation OpenRouter, les modèles chinois ont représenté 34 250 milliards de jetons consommés sur la semaine du 3 au 9 août 2026, dépassant les modèles américains pour la quinzième

semaine consécutive. Les estimations de part varient fortement selon les périodes et les périmètres retenus (de 30 % à plus de 60 % selon les publications) : ce chiffre doit être manié avec précaution.

MIT Technology Review rapportait le 12 février 2026 que les dérivés de Qwen représentaient plus de 40 % des nouvelles variantes de modèles de langage publiées sur Hugging Face au 4 août 2025, contre environ 15 % pour les dérivés de Llama.

IV.2 Les modèles chinois de référence (2026)

Modèle	Laboratoire	Date	Taille	Licence
DeepSeek V4-Pro	DeepSeek	24 avril 2026	1 600 Md param. (49 Md actifs)	MIT
DeepSeek V4-Flash	DeepSeek	24 avril 2026	284 Md param. (13 Md actifs)	MIT
Qwen 3.5	Alibaba	Février 2026	Famille	Apache 2.0
Qwen 3.8-Max	Alibaba	2026	Non publié	Aucun dépôt public
Kimi K2.5	Moonshot AI	Février 2026	—	« modified-MIT »
Kimi K3	Moonshot AI	16 juillet 2026 (poids : 27 juillet)	2 800 Md param.	« kimi-k3 »
GLM-5.2	Zai (ex-Zhipu)	2026	—	MIT
GLM-5.3	Zai	25 août 2026	—	« glm-5.3 » (propriétaire)
MiniMax M3	MiniMax	2026	—	« minimax-community »

DeepSeek V4, publié le 24 avril 2026 sous licence MIT, dispose d’une fenêtre de contexte d’un million de jetons et est tarifé 1,74 dollar par million de jetons en entrée et 3,48 dollars en sortie, contre 5 et 30 dollars pour GPT-5.5 et 5 et 25 dollars pour Opus 4.7 à la même date. DeepSeek affirme que V4-Pro n’accuse qu’un retard de trois à six mois sur les meilleurs modèles fermés.

Kimi K3, publié par Moonshot AI le 16 juillet 2026 avec ouverture des poids le 27 juillet, revendique 2 800 milliards de paramètres, soit le plus grand modèle à poids ouverts publié à cette date. Il se classe premier sur le classement Arena.ai Frontend Code, devant Claude Fable 5 et GPT-5.6 Sol, et troisième sur GDPval-AA v2.

IV.3 La politique publique chinoise

Le Conseil des affaires de l’État chinois a publié en août 2024 un projet de politique encourageant les universités à valoriser les contributions open source dans l’évaluation académique. La Chine impose depuis

2025 un étiquetage obligatoire des contenus générés par IA et un enregistrement préalable des modèles génératifs auprès de l'administration du cyberspace.

Lors de la réunion ministérielle de l'APEC à Chengdu, dont la déclaration a été publiée le 26 juillet 2026, les vingt et une économies membres, y compris les États-Unis et la Chine, ont affirmé soutenir « les modèles et projets open source qui mettent en œuvre une assurance de sécurité renforcée » et reconnu « le rôle important des approches open source de confiance ».

La commission américaine US-China Economic and Security Review a qualifié en mars 2026 la position des modèles chinois sur Hugging Face et OpenRouter d'« avantage compétitif auto-renforçant ».

V. Le repli de l'ouverture : acteurs américains et signaux chinois

V.1 Meta

Meta a publié en avril 2026 Muse Spark, développé par Meta Superintelligence Labs : un modèle propriétaire, accessible uniquement par interface applicative en nuage, sans poids téléchargeables. La société indique que les modèles Llama existants « resteront disponibles en open source », sans engagement sur les générations futures. Le cumul historique des téléchargements de Llama est estimé à 1,2 milliard. Andrew Ng a qualifié ce virage de « perte significative » pour l'écosystème des poids ouverts.

Les licences Llama 4 (Scout et Maverick) ne sont pas approuvées par l'Open Source Initiative : elles comportent une validation manuelle et un seuil de 700 millions d'utilisateurs actifs mensuels déclenchant des obligations de marque et de dénomination des dérivés.

Meta avait par ailleurs refusé, le 18 juillet 2025, de signer le code de bonnes pratiques européen pour les modèles à usage général, alors qu'OpenAI et Anthropic l'ont signé.

V.2 OpenAI

OpenAI a publié le 5 août 2025 gpt-oss-120b (116,8 milliards de paramètres, 5,1 milliards actifs) et gpt-oss-20b (21 milliards de paramètres, 3,6 milliards actifs) sous licence Apache 2.0 : ses premiers modèles à poids ouverts depuis GPT-2 en 2019. Des variantes gpt-oss-safeguard ont suivi en octobre 2025. À la mi-2026, aucun modèle de base successeur n'avait été publié.

Le plan « Winning the Race: America's AI Action Plan », publié par la Maison-Blanche le 23 juillet 2025, comporte une section intitulée « Encourage Open Source and Open Weight AI » et recommande l'exportation de « paquets complets » de la pile technologique américaine.

V.3 Mistral AI

Mistral AI a publié le 2 décembre 2025 la famille Mistral 3 sous licence Apache 2.0 : Mistral Large 3 (675 milliards de paramètres, 41 milliards actifs), Ministral 3 en versions 14B, 8B et 3B. L'entreprise affirme que « l'avenir de l'IA est ouvert ».

Le portefeuille n'est cependant pas intégralement ouvert. Plusieurs modèles — Mistral Medium 3.5, Codestral, Magistral Medium, Pixtral selon les versions — relèvent de licences restrictives ou d'un accès par interface applicative uniquement. Les sources consultées divergent sur le statut exact de certaines de ces références : ce point demande vérification directe auprès de la documentation de l'éditeur.

V.4 Les signaux de fermeture côté chinois

Le 18 août 2026, Z.ai a annoncé reporter de deux semaines la publication des poids de GLM-5.3, indiquant tester et renforcer ses contrôles de sécurité avant mise à disposition large. Le modèle atteignait 84,5 % sur le référentiel CyberGym, devant ses concurrents fermés, et avait identifié 2 436 vulnérabilités sur 269 projets, dont 1 097 de sévérité moyenne à critique. C'est, à la connaissance des sources consultées, le premier cas documenté de retenue volontaire d'un laboratoire chinois pour motif de sécurité.

Les poids ont été publiés le 25 août 2026, mais sous une licence propriétaire « glm-5.3 » et non plus sous MIT : les opérateurs de modèle en tant que service dont le chiffre d'affaires cumulé sur douze mois glissants dépasse 10 milliards de dollars doivent obtenir une revue de sécurité de Z.ai avant usage commercial.

Un audit de licences publié le 16 août 2026 documente la diffusion de seuils commerciaux dans les licences chinoises : Kimi K2.5 et K3 imposent des obligations de marque au-delà de 100 millions d'utilisateurs actifs mensuels ou 20 millions de dollars de revenu mensuel, et un accord séparé au-delà de 20 millions de dollars de revenu sur douze mois glissants ; MiniMax M3 impose une autorisation écrite au-delà de 20 millions de dollars annuels, assortie de dix-neuf catégories d'usages prohibés. Par ailleurs, le modèle haut de gamme d'Alibaba, Qwen 3.8-Max, ne dispose d'aucun dépôt public.

V.5 Le rachat de Hugging Face par Nvidia

Le 26 août 2026, The Information a révélé les discussions avancées entre Nvidia et Hugging Face. L'accord définitif a été annoncé le 3 septembre 2026, pour un montant de 12,93 milliards de dollars. Hugging Face, cofondée par des entrepreneurs français, avait été valorisée 4,5 milliards de dollars lors de son tour de 2023 et réalisait environ 150 millions de dollars de chiffre d'affaires annuel. La société avait refusé fin 2025 un investissement de 500 millions de dollars de Nvidia sur une valorisation de 7 milliards.

Jensen Huang a déclaré que la plateforme « resterait ouverte à l'ensemble de l'écosystème de l'IA ». L'opération doit passer un examen antitrust. Les commentaires relevés soulignent que l'annonce ne précise pas qui détiendra les droits de décision sur les contenus hébergés, ni ce qui se produirait si la politique américaine exigeait une modification de la disponibilité des modèles chinois.

VI. La géopolitique du calcul

VI.1 Répartition mondiale

Selon Epoch AI, les États-Unis concentraient en mai 2025 environ 75 % de la performance mondiale des grappes de calcul GPU, la Chine environ 15 %, le reste du monde — Allemagne, Japon, France compris — environ 10 %. Epoch relève que les puissances traditionnelles du calcul intensif « ne jouent plus qu'un rôle marginal » dans le paysage des grappes d'IA.

Le rapport AI Index 2026 de Stanford HAI, publié le 13 avril 2026, établit les ordres de grandeur suivants : capacité de calcul mondiale dédiée à l'IA de 17,1 millions d'équivalents H100, soit une multiplication par trente depuis 2021 ; capacité électrique des centres de données d'IA de 29,6 gigawatts, équivalente à la pointe de consommation de New York ; investissement privé américain dans l'IA de 285,9 milliards de dollars contre 12,4 milliards pour la Chine, soit un rapport de 23,1 (hors fonds d'orientation publics chinois, estimés à environ 184 milliards) ; écart de performance entre meilleurs modèles américains et chinois réduit à 2,7 %, contre 17,5 à 31,6 points en mai 2023 ; indice de transparence des modèles de fondation en recul à 40 points contre 58 ; 362 incidents liés à l'IA documentés en 2025, contre 233 en 2024 ; adoption organisationnelle de l'IA à 88 %.

Nvidia a déclaré un chiffre d'affaires « IA souveraine » de 30 milliards de dollars sur son exercice 2026, triplé en un an, soit environ 1 % de ses 215,9 milliards de dollars de ventes. Les marchés cités sont le Canada, la France, les Pays-Bas, Singapour, le Royaume-Uni et l'Inde.

VI.2 Contrôles américains à l'exportation

Date	Mesure
Octobre 2022	Premiers contrôles à l'exportation de semi-conducteurs vers la Chine
Avril 2025	Licence exigée pour les H20 vers la Chine
Mai 2025	Abrogation de l'« AI Diffusion Rule »
Décembre 2025	Annnonce d'un accord de vente H200 / MI325X avec prélèvement de 25 %
14 janvier 2026	Proclamation présidentielle instaurant un droit de douane section 232
15 janvier 2026	Règle BIS : H200 et MI325X passent de la présomption de refus à l'examen au cas par cas
21 janvier 2026	AI OVERWATCH Act (H.R. 6875) adopté en commission des affaires étrangères par 42 voix contre 2
Février 2026	Le Département d'État suspend les expéditions de H200 dans l'attente d'une revue de sécurité

Le plafond d'exportation autorise environ 850 000 équivalents H200 vers la Chine, dans la limite de 50 % des ventes américaines cumulées. Les accélérateurs de classe Blackwell (B200, GB200, GB300) restent sous présomption de refus. Le Conseil des relations étrangères américain a qualifié cette politique de « stratégiquement incohérente et inapplicable ».

VI.3 Capacités chinoises domestiques

Les accélérateurs Huawei Ascend équiperont environ 45 % des centres de données chinois et généreront 15 milliards de dollars de chiffre d'affaires annuel, en croissance de 30 %. Leur fabrication dépend de SMIC, plafonné à une gravure 7 nanomètres avec des rendements inférieurs à 50 %, en raison notamment du refus d'ASML de fournir des équipements de lithographie avancés. Un Ascend 960 est annoncé pour fin 2026. Ces chiffres proviennent de la presse spécialisée et non de sources primaires : ils doivent être considérés comme des estimations.

VI.4 Capacités européennes

Le processeur européen Rhea1, conçu par SiPearl dans le cadre de la European Processor Initiative lancée en 2019, est sorti des chaînes de TSMC et se trouvait en phase de validation matérielle et logicielle en mai 2026. Il comporte 80 cœurs Arm Neoverse V1 et 64 gigaoctets de mémoire HBM pour environ 350 watts. Il doit équiper une extension du supercalculateur Jupiter (Jülich) fin 2026, à raison de 1 300 serveurs biprocesseurs. Le successeur Rhea2 est destiné au calculateur exascale français Alice Recoque. Point essentiel : Rhea1 est un processeur

généraliste, non un accélérateur d’entraînement. Il ne constitue donc pas une alternative aux accélérateurs Nvidia pour l’entraînement de modèles de fondation.

VII. Le cadre européen

VII.1 Le règlement sur l’intelligence artificielle

Calendrier d’application du règlement (UE) 2024/1689, tel que modifié par l’omnibus numérique adopté par le Parlement européen le 16 juin 2026 (423 voix pour, 57 contre, 174 abstentions) et par le Conseil le 29 juin 2026 :

Échéance	Obligations
2 février 2025	Interdictions de pratiques ; obligation de littératie en IA
2 août 2025	Obligations des fournisseurs de modèles à usage général
2 août 2026	Article 50 (transparence) ; autorités nationales compétentes ; sanctions ; bacs à sable réglementaires
2 décembre 2026	Marquage des contenus déjà commercialisés ; interdiction des systèmes de contenus intimes non consentis
2 décembre 2027	Systèmes à haut risque de l’annexe III (report depuis le 2 août 2026)
2 août 2028	Systèmes à haut risque de l’annexe I (report depuis le 2 août 2027)

Les sanctions au titre de l’article 50 peuvent atteindre 15 millions d’euros ou 3 % du chiffre d’affaires mondial annuel.

L’exemption open source figure à l’article 53, paragraphe 2 : les fournisseurs de modèles à usage général publiés sous licence libre et ouverte sont dispensés des obligations de documentation technique (annexe XI) et d’information des fournisseurs en aval (annexe XII), sous quatre conditions cumulatives : licence libre et ouverte ; licence autorisant l’accès, l’usage, la modification et la distribution ; poids publiquement disponibles ; architecture et informations d’usage publiquement disponibles. Cette exemption ne s’applique pas aux modèles présentant un risque systémique (seuil indicatif de 10^25 opérations en virgule flottante à l’entraînement), qui restent soumis à l’intégralité de l’article 53 et à l’article 55. Les obligations relatives au droit d’auteur (article 53, paragraphe 1, point c) et au résumé public des données d’entraînement (point d) s’appliquent en toute hypothèse.

Le Comité européen de la protection des données et le Contrôleur européen de la protection des données ont rendu le 11 février 2026 un avis conjoint sur l’omnibus numérique. Ils s’opposent fermement à la modification de la définition des données à caractère personnel, qui « va bien au-delà d’une modification ciblée ou technique ». Anu Talus, présidente du CEPD : « La simplification est essentielle pour réduire la charge administrative et renforcer la compétitivité de l’Union — mais pas au détriment des droits fondamentaux. »

VII.2 Le plan « AI Continent » et l’investissement

Le plan d’action « AI Continent », publié le 9 avril 2025, repose sur cinq piliers : infrastructure de calcul (19 usines d’IA et jusqu’à cinq gigafactories), accès aux données de qualité, adoption sectorielle, compétences et talents, mise en œuvre du règlement sur l’IA. L’initiative InvestAI mobilise 200 milliards d’euros, dont 20 milliards pour les gigafactories. La Commission vise un triplement de la capacité de centres de données de l’Union en cinq à sept ans. Le plan relève que 13,5 % des entreprises européennes utilisaient alors l’IA.

Les stratégies « Apply AI » et « AI in Science » ont été présentées le 8 octobre 2025. La première mobilise environ 1 milliard d’euros pour l’adoption sectorielle (santé, énergie, mobilité, industrie manufacturière, construction, agroalimentaire, défense, communications, culture) et promeut un principe d’« AI first ». La seconde crée l’institut virtuel RAISE, avec 58 millions d’euros pour les réseaux de talents et 600 millions au titre d’Horizon Europe pour l’accès au calcul.

L’entreprise commune EuroHPC a lancé le 30 juillet 2026 l’appel à candidatures pour les gigafactories d’IA : jusqu’à sept installations réparties dans au moins sept États membres, avec possibilité de consortiums transfrontaliers ; date limite de dépôt le 12 novembre 2026 ; sélection attendue début 2027 ; mise en service dans les dix-huit mois suivant la sélection ; plus de 20 milliards d’euros d’investissement privé attendus.

En France, le consortium AION, mené par Scaleway (groupe iliad) avec OpCore, GENCI, Inria, VSORA, Kyutai, Sopra Steria, SiPearl, Artefact, Eviden Bull, ZML, Hugging Face et H Company, a officialisé une candidature portant sur 288 000 équivalents H100 et 200 mégawatts.

Le réseau des usines d’IA d’EuroHPC comptait treize sites après la deuxième vague de sélection de mars 2025, dont l’usine française AI2F pilotée par GENCI avec Inria, le CNRS, le CEA, le CINES, France Universités, la Mission French Tech et le Hub France IA. Les six projets de la deuxième vague représentent environ 485 millions d’euros d’investissement national et européen.

VII.3 Le règlement sur le cloud et l’IA (CADA)

La Commission européenne a présenté le 3 juin 2026 le paquet « souveraineté technologique européenne », dont la pièce maîtresse est la proposition de règlement sur le développement du cloud et de l’IA (Cloud and AI Development Act, CADA). Le texte établit quatre niveaux d’assurance de souveraineté pour les services en nuage :

Niveau	Exigence
1	Données traitées et stockées dans l’Union
2	Indépendance du fournisseur à l’égard des pays tiers ; chaîne logicielle transparente ; absence d’interrupteur à distance
3	Fournisseur non contrôlé par un pays tiers ; exigences complémentaires (nationalité du personnel)
4	Composants non contrôlés par un pays tiers ; certification de cybersécurité du plus haut niveau

Les estimations d’impact avancées dans l’analyse publiée par Lawfare indiquent qu’environ 70 % des marchés relèveraient du niveau 1 — accessible aux filiales européennes des fournisseurs américains —, environ 20 % du niveau 2, moins de 10 % du niveau 3 et environ 1 % du niveau 4. OVHcloud estime avoir besoin que 15 % de la commande publique soit réservée pour atteindre une échelle compétitive. Trois fournisseurs américains (Microsoft, Amazon Web Services, Google) détiennent plus de 70 % du marché européen du cloud, les fournisseurs européens environ 15 %. Des associations professionnelles américaines, canadiennes, japonaises et australiennes ont demandé une révision du texte au nom de la non-discrimination et de l’ouverture aux partenaires commerciaux.

Le 20 avril 2026, la Commission européenne a attribué un marché cloud de 180 millions d’euros sur six ans à OVHcloud (avec Clever Cloud et DEEP de POST Luxembourg), Scaleway, S3NS (Thales et Google Cloud, avec Proximus) et StackIT, dans le cadre du système d’acquisition dynamique « Cloud III ».

VIII. Les positions institutionnelles françaises

VIII.1 La commission d'enquête de l'Assemblée nationale

La commission d'enquête sur les dépendances structurelles et les vulnérabilités systémiques dans le secteur du numérique et les risques pour l'indépendance de la France a été créée le 3 février 2026. Présidée par Philippe Latombe (Les Démocrates, Vendée), avec Cyrielle Chatelain (Écologiste et social, Isère) comme rapporteure, elle a adopté son rapport le 8 juillet 2026 et l'a rendu public le 15 juillet 2026, au terme de 45 auditions ayant entendu près de 112 personnes, et a formulé 47 recommandations.

Constats et recommandations documentés :

- Les géants technologiques américains capteraient 83 % des dépenses cloud européennes, soit environ 264 milliards d'euros par an.
- Des données françaises de défense et d'éducation sont hébergées à l'étranger, ainsi que des données de santé confiées à Microsoft malgré les réserves de la CNIL.
- La commission recommande une action spécifique (« golden share ») de l'État dans Mistral AI et ChapsVision : une participation de 5 % assortie de droits spécifiques permettant de bloquer l'entrée d'actionnaires étrangers sans faire obstacle aux levées de fonds. Le fondement invoqué est l'article 31-1 de l'ordonnance du 20 août 2014, qui permet à l'État des droits particuliers indépendamment de sa part de capital, dont un droit d'opposition aux cessions d'actifs stratégiques. La mise en œuvre supposerait une augmentation de la participation de Bpifrance.
- Un moratoire sur la construction de centres de données étrangers en France est également proposé.

Vincent Strubel, directeur général de l'ANSSI, auditionné par la commission : « Si demain, pendant six mois, on est privés d'accès à la technologie américaine, privés des mises à jour des logiciels américains, on a un énorme problème. »

Arthur Mensch, auditionné le 12 mai 2026, a estimé que l'IA pourrait absorber 10 % de la masse salariale européenne, soit environ 1 000 milliards d'euros par an, représentant autant de déficit commercial si la technologie est importée. Il a alerté sur le risque de captation du surplus électrique français de 9 gigawatts par des centres de données américains et critiqué la fragmentation de la mise en œuvre du

règlement sur l'IA entre vingt-sept autorités nationales : « Réguler pour se défendre, en général, ça ne marche pas. Cela favorise les plus gros acteurs. »

VIII.2 La Cour des comptes

Dans un rapport publié le 19 novembre 2025, la Cour des comptes alerte sur le retard pris par la transformation des services publics en matière d'IA, qu'elle juge « très modeste ». Les feuilles de route ministérielles lancées en février 2025 n'avaient pas produit de résultats substantiels à leur échéance de juin. La DINUM estime à environ 100 millions d'euros par an, à compter de 2026, le coût du déploiement d'infrastructures et de services d'IA à l'échelle de l'État — financement non alloué. La Cour recommande la création d'un secrétariat général à l'IA rattaché au Premier ministre, l'utilisation de la commande publique au profit des solutions françaises, et s'appuie sur les start-up nationales et les solutions open source pour réduire la dépendance aux fournisseurs extérieurs. Elle cite l'estimation selon laquelle au moins 38 % des emplois publics dans le monde seront affectés par l'IA.

La Cour a par ailleurs publié le 8 janvier 2026 des observations définitives sur France Travail et l'intelligence artificielle, soulignant des résultats opérationnels contrastés après dix ans d'usage.

VIII.3 La mission IGF-IGA-IGAS

Le rapport « Le déploiement de l'intelligence artificielle dans les administrations publiques : analyse comparée, enjeux et conditions de réussite » (avril 2026, mis en ligne en juillet 2026), conduit par Jérôme Letier, Arnaud Mercier et Pierre Cunéo, établit un état des lieux comparé avec treize administrations étrangères. Principaux constats :

- Environ 40 % des agents publics utiliseraient l'IA de manière informelle (« IA clandestine »), avec des risques de sécurité des données et d'équité.
- L'inventaire de l'administration d'État recense 141 produits d'IA dans 80 entités publiques, avec un déploiement très inégal (de 0 à 74 projets selon le ministère).
- 90 % des métropoles sont engagées, contre 73 % des communes de moins de 3 500 habitants sans aucun projet d'IA.
- 13 à 20 % des agents publics sont significativement exposés à l'IA selon la méthodologie de l'Organisation internationale du travail.
- Gains documentés : délai moyen de traitement documentaire ramené de 19 à 3 jours ; taux d'abandon d'appels dans des agents conversationnels municipaux ramené de 55 % à 8 % ; économies de temps sur les fonctions support estimées à 20-40 %.

- Dépenses 2023–2025 : 8,7 millions d’euros pour la DINUM, 26,2 millions pour les ministères, 18,7 millions pour France Travail. Les coûts « dépassent souvent les prévisions ».
- La qualification SecNumCloud représente un surcoût de 20 à 40 %.
- La mission privilégiée « la maîtrise des infrastructures et des données plutôt que l’autarcie technologique », recommande des modèles frugaux, des clauses contractuelles types garantissant la réversibilité et l’autonomie de l’acheteur, et une trajectoire pluriannuelle d’infrastructures mutualisées (processeurs graphiques, hébergement).

VIII.4 La DINUM et le plan « Notre IA »

Le 16 juin 2026, la DINUM a présenté le plan « Notre IA » pour les services publics. Quatre outils souverains sont déployés : L’Assistant (agent conversationnel développé avec Mistral AI), Transcript (transcription audio), DiplolA (traduction en 64 langues) et Visio (avec comptes rendus de réunion). L’Assistant a été testé pendant dix mois auprès de 10 000 agents avant sa généralisation. L’ensemble est hébergé sur l’infrastructure Outscale qualifiée SecNumCloud. Les objectifs affichés sont l’élimination de l’« IA clandestine », des gains de temps allant jusqu’à 16 % sur la synthèse documentaire, et une relocalisation complète des compétences numériques d’ici 2027.

VIII.5 La CNIL

La CNIL a publié le 7 avril 2026 son programme de travail 2026, dont les priorités relatives à l’IA comprennent : la finalisation des recommandations sur l’IA dans le travail et la santé ; l’analyse automatique de la voix ; des outils d’accompagnement pour les délégués à la protection des données ; la clarification des conséquences de l’application du RGPD aux modèles d’IA et de la répartition des rôles entre concepteurs et déployeurs ; la préparation de son rôle d’autorité de surveillance du marché au titre du règlement sur l’IA.

VIII.6 Le Sénat et la commande publique

En réponse à une question écrite de la sénatrice Cathy Apourceau-Poly (publiée le 12 mars 2026, réponse publiée le 28 mai 2026) portant sur la souveraineté technologique et l’intégrité des données des citoyens français face à l’IA, le Gouvernement a indiqué : la commande publique en matière de cloud a représenté 84 millions d’euros en 2025, dont 69 % vers des opérateurs européens et 22 millions d’euros d’offres qualifiées SecNumCloud ; la qualification SecNumCloud exige un siège social dans l’Union, un hébergement dans l’Union et une participation individuelle d’actionnaire non européen plafonnée à 24 %.

VIII.7 Le Health Data Hub

Le 10 février 2026, les ministères de la Santé, de la Réforme de l'État et du Numérique ont annoncé que Microsoft n'hébergerait plus le Health Data Hub. Un appel d'offres devait désigner d'ici fin mars 2026 un prestataire qualifié SecNumCloud, protégé des lois à portée extraterritoriale. La plateforme, lancée en 2019, fonctionnait au ralenti en raison des réserves répétées de la CNIL sur les transferts de données. En avril 2025, la ministre déléguée au Numérique Clara Chappaz avait déjà appelé à une migration vers un opérateur sécurisé.

IX. Les sommets internationaux

IX.1 Paris, février 2025

Le Sommet pour l'action sur l'IA s'est tenu les 10 et 11 février 2025 au Grand Palais, coprésidé par Emmanuel Macron et Narendra Modi. Annonces principales : 109 milliards d'euros d'investissements privés en France, dont 30 à 50 milliards d'origine émiratie pour un campus de centres de données et 20 milliards de la société canadienne Brookfield ; lancement de l'initiative européenne InvestAI dotée de 200 milliards d'euros, dont 20 milliards pour quatre gigafactories ; création de la fondation Current AI, dotée de 400 millions, soutenue par neuf gouvernements (Finlande, France, Allemagne, Chili, Inde, Kenya, Maroc, Nigeria, Slovénie). Cinquante-huit pays ont signé la déclaration sur une IA inclusive et durable, dont la France, la Chine et l'Inde ; les États-Unis et le Royaume-Uni ont refusé de la signer.

Le 6 février 2025, le comité interministériel avait présenté la troisième phase de la stratégie nationale française : 360 millions d'euros pour neuf IA-Clusters visant 100 000 personnes formées d'ici 2030, 35 sites de centres de données identifiés, feuilles de route ministérielles, création de l'INESIA (institut d'évaluation de l'IA), objectif de sensibilisation de 2 millions de Français d'ici fin 2027.

IX.2 New Delhi, février 2026

Le sommet AI Impact s'est tenu à New Delhi du 17 au 23 février 2026. La déclaration de New Delhi a été adoptée par 88 pays et organisations internationales, États-Unis et Chine compris — ce qui constitue un changement par rapport à Paris. Les engagements d'investissement annoncés totalisent environ 200 milliards de dollars : 110 milliards par Reliance Industries/Jio sur sept ans, 100 milliards par Adani pour des centres de données alimentés en énergies renouvelables d'ici 2035, 50 milliards par Microsoft dans les pays du Sud d'ici 2030.

Thèmes retenus : démocratisation de l'accès, santé et éducation, IA open source et souveraine, concentration du calcul (les États-Unis et la Chine contrôlant environ 90 % de l'infrastructure mondiale), fiabilité des modèles pour les langues sous-représentées.

L'Inde a dévoilé le 19 février 2026 trois modèles souverains : deux modèles Sarvam de 30 et 105 milliards de paramètres pour les langues indiennes ; Param2, modèle de mélange d'experts de 17 milliards de paramètres de l'initiative BharatGen ; Vachana TTS de Gnani.ai, couvrant douze langues indiennes avec clonage vocal à partir de moins de dix secondes d'enregistrement.

Michael Kratsios (Maison-Blanche) : « La véritable souveraineté en IA, c'est posséder la meilleure technologie de sa catégorie. » Arthur Mensch y a défendu l'idée que toute organisation doit disposer du « bouton marche-arrêt ».

X. La demande : adoption en France et dans les territoires

X.1 Entreprises

Selon l'Insee (« Les technologies de l'information et de la communication dans les entreprises en 2025 », publié le 21 juillet 2026), 18 % des entreprises implantées en France déclaraient utiliser au moins une technologie d'intelligence artificielle en 2025, contre 10 % en 2024 et 6 % en 2023. La répartition par taille est très contrastée : 15 % pour les entreprises de 10 à 49 salariés, 31 % pour celles de 50 à 249 salariés, 58 % pour celles de 250 salariés et plus. La moyenne de l'Union européenne se situe à 20 %. Parmi les entreprises utilisatrices, 56 % recourent à l'analyse de texte, 43 % à la génération de texte ou de parole et 41 % à la génération d'image ou de vidéo.

Une enquête Bpifrance publiée en mars 2026 indique que 47 % des PME françaises ont lancé au moins un projet d'IA, contre 29 % début 2024, et que le taux atteint 95 % pour les entreprises de plus de 200 salariés.

Freins déclarés : difficulté de recrutement et de fidélisation des compétences (60 %), cadre réglementaire perçu comme flou (52 %), absence de stratégie préalable (43 %), biais algorithmiques et enjeux éthiques (39 %), intégration aux systèmes existants (32 %).

L'investissement moyen d'un premier projet est estimé à 79 000 euros, souvent doublé par les coûts cachés de préparation des données et de formation.

X.2 Collectivités territoriales

Le baromètre 2025 de l'Observatoire Data Publica (publié le 14 novembre 2025, 292 collectivités interrogées de mai à juillet 2025) établit que 84 % des collectivités de plus de 3 500 habitants utilisent l'IA générative, contre 52 % en 2024 ; le taux atteint 91 % pour les départements. Les usages portent d'abord sur la gestion interne (45 %, contre 29 % en 2024) puis sur la relation à l'utilisateur (20 %). Freins : manque de temps et de compétences (62 %), confiance limitée dans les systèmes (46 %), qualité insuffisante des données (41 %, contre 26 % en 2024). Point saillant pour la souveraineté : 82 % des collectivités placent la localisation des données en France ou en Europe parmi leurs critères, et 80 % de celles qui expérimentent ont mis en place des mesures d'IA de confiance, dont des chartes éthiques (57 %).

X.3 Éducation

Le ministère de l'Éducation nationale a annoncé le 7 février 2025 une série de mesures : publication d'une charte d'usage de l'IA au printemps 2025 ; développement, avec 20 millions d'euros de France 2030, d'un outil souverain d'assistance aux enseignants (préparation de cours, évaluation) disponible à la rentrée 2026-2027 ; formation Pix obligatoire pour les élèves de quatrième et de seconde à partir de la rentrée 2025 ; généralisation d'un assistant conversationnel pour la gestion des ressources humaines de plus de 1,2 million de personnels. Le taux d'adoption de l'IA par les enseignants était alors estimé à moins de 20 %.

X.4 Énergie et raccordement

Selon RTE (données publiées en mai 2026), les centres de données consomment environ 10 térawattheures par an en France, soit environ 2 % de la consommation nationale, répartis sur environ 300 installations. Environ 18 gigawatts de capacité étaient réservés en mai 2026 au titre d'environ 80 projets, contre environ 5 gigawatts pour 40 projets fin 2024. Les projections RTE retiennent 15 à 20 térawattheures en 2030 (environ 3 % de la consommation) et 23 à 28 térawattheures en 2035 (environ 4 %). RTE souligne que les centres de données existants ne consomment en moyenne que 20 % de la puissance demandée et que la montée en charge s'étale sur dix à quinze ans.

XI. Tableau comparatif des dépendances par couche

Couche	Acteurs dominants	Position européenne	Position française
Accélérateurs d'entraînement	Nvidia, AMD ; Huawei en Chine	Marginale ; SiPearl produit un CPU, pas un accélérateur	Nulle en propre ; VSORA au stade du projet
Lithographie	ASML (Pays-Bas)	Dominante — unique point de force structurel	Indirecte (fournisseurs)
Centres de données et énergie	Hyperscalers américains ; opérateurs régionaux	En construction (gigafactories)	Atout électrique ; ~18 GW de demandes de raccordement
Cloud	Microsoft, AWS, Google (> 70 % du marché UE)	~15 % pour les acteurs européens	OVHcloud, Scaleway, Outscale, Cloud Temple
Modèles fermés de frontière	OpenAI, Anthropic, Google	Mistral AI au second rang mondial	Mistral AI
Modèles à poids ouverts	Alibaba, DeepSeek, Moonshot, Zai, MiniMax	Mistral AI (Apache 2.0)	Mistral AI, Kyutai
Plateformes de distribution	Hugging Face (rachetée par Nvidia en septembre 2026)	Fondateurs français, propriété américaine	—
Normes et droit	Union européenne (règlement sur l'IA)	Dominante	Contributrice

XII. Sources et bibliographie sélective

Sources primaires officielles. Commission européenne, plan d'action « AI Continent » (9 avril 2025) ; stratégies « Apply AI » et « AI in Science » (8 octobre 2025) ; proposition de règlement sur le développement du cloud et de l'IA (3 juin 2026). EuroHPC, appel « AI Gigafactories » (30 juillet 2026) ; sélection des usines d'IA (mars 2025). Règlement (UE) 2024/1689 et omnibus numérique (Parlement européen, 16 juin 2026 ; Conseil, 29 juin 2026). CEPD et CEPD-EDPS, avis conjoint sur l'omnibus numérique (11 février 2026). Assemblée nationale, commission d'enquête sur les dépendances structurelles et les vulnérabilités systémiques dans le secteur du numérique (créée le 3 février 2026, rapport adopté le 8 juillet 2026, publié le 15 juillet 2026). Cour des comptes, rapport sur l'IA dans les services publics (19 novembre 2025) ; observations définitives France Travail et l'IA (8 janvier 2026). IGF-IGA-IGAS, « Le déploiement de l'intelligence artificielle dans les administrations publiques » (avril 2026). DINUM, plan « Notre IA » (16 juin 2026). CNIL, programme de travail 2026

(7 avril 2026). Sénat, question écrite n° 260308036 et réponse du Gouvernement (12 mars et 28 mai 2026). Ministère de l'Éducation nationale, mesures IA et éducation (7 février 2025). Insee, Insee Première n° 2120 (21 juillet 2026). RTE, « Les data centers en chiffres clés » (mai 2026). Bureau of Industry and Security, règle du 15 janvier 2026 (Federal Register 2026-00789). Maison-Blanche, « America's AI Action Plan » (23 juillet 2025). Ministère indien des Affaires étrangères, déclaration de New Delhi (23 février 2026).

Publications d'entreprises. Mistral AI, annonce de la série D (8 septembre 2026) ; annonce de la série C (9 septembre 2025) ; « Introducing Mistral 3 » (2 décembre 2025) ; AI Now Summit (28 mai 2026) ; Regional Endpoints et hébergement de GLM-5.2 (11 et 14 août 2026). MGX, Campus AI (20 mai 2025 et 1er juin 2026). Hugging Face, « State of Open Models: Summer 2026 » (14 août 2026). OpenAI, « Introducing gpt-oss » (5 août 2025).

Sources académiques et instituts. Stanford HAI, AI Index Report 2026 (13 avril 2026). Epoch AI, « Trends in AI supercomputers » et données de répartition par pays (mai 2025). Institut Montaigne, François Godement, « Offensive chinoise en Europe : l'open source, nouveau cheval de Troie numérique » (7 septembre 2026). Chaire numérique, gouvernance et souveraineté de Sciences Po, rapport « Buy European Tech Act » (2024). US-China Economic and Security Review Commission (mars 2026).

Presse spécialisée et générale. TechCrunch, « Mistral raises €3B as sovereign AI becomes big business » (8 septembre 2026) et « Nvidia closes in on Hugging Face acquisition » (26 août 2026). Bloomberg (8 septembre 2026 et 15 août 2026). CNBC (8 septembre 2026). MIT Technology Review, « What's next for Chinese open-source AI » (12 février 2026). Le MagIT (9 septembre 2025, 27 mai 2026, 20 avril 2026). South China Morning Post (9 avril 2026 et 21 août 2026). VentureBeat, Kimi K3 (16 juillet 2026). DataCamp, DeepSeek V4 (24 avril 2026). Datacenter Dynamics (20 mai 2025). Acteurs publics (19 novembre 2025). Clubic (septembre 2026). LCP (8 juillet 2026). Techzine (14 août 2026). Developpez.com (18 août 2026).