Au-delà du token suivant : trois architectures pour trois manques du paradigme LLM

World models, mémoire et JEPA comme réponses partielles à la dynamique, à la persistance et à l'espace de prédiction

Note technique distinguant trois familles d'architectures (world models, JEPA, modèles à mémoire) comme réponses à trois angles morts du paradigme LLM

1. Le manque fondateur

Le paradigme dominant de l'intelligence artificielle en 2026 reste le modèle auto-régressif entraîné par maximum de vraisemblance sur un corpus de symboles. Sa tâche fondamentale est simple à énoncer : prédire le prochain élément d'une séquence étant donné les éléments précédents. La réussite empirique de cette tâche est considérable. Elle a produit des systèmes capables de résumer, traduire, programmer, raisonner localement, dialoguer, reformuler des connaissances et manipuler des structures symboliques avec une efficacité qui aurait paru invraisemblable dix ans plus tôt.

Mais cette réussite ne doit pas être mal nommée. Ce qui passe, dans un grand modèle de langage, pour de la compréhension, de la mémoire ou de la simulation du monde n'est pas directement optimisé comme tel. C'est une propriété émergente d'un apprentissage séquentiel sur des traces symboliques. Le modèle apprend à produire des continuations plausibles dans l'espace des tokens. Il n'apprend pas, par cette seule opération, un modèle explicite de la dynamique du monde, une mémoire biographique persistante[S1], ni un espace de prédiction adapté à des phénomènes non linguistiques.

Le diagnostic architectural part de là. Le modèle auto-régressif standard présente trois angles morts distincts.

1. Le premier est celui de la dynamique. Le modèle sait estimer ce qui est susceptible de venir après une séquence ; il ne sait pas, par construction, ce qui se passerait si une action était réalisée dans un environnement. La différence entre prédiction passive et prédiction conditionnée à l'action est fondamentale pour la planification, la robotique, la cognition incarnée, la simulation clinique ou tout système soumis à des interventions. Un LLM peut décrire une conséquence ; il n'a pas nécessairement appris la dynamique causale qui la produit.

2. Le deuxième est celui de la persistance. La fenêtre de contexte est un dispositif de calcul, pas une mémoire au sens fort. Elle peut être très longue, parfois jusqu'au million de tokens, mais elle reste essentiellement plate : elle ne distingue pas nativement information de référence, souvenir daté, état courant, instruction transitoire, préférence stable ou événement singulier. Elle ne persiste pas naturellement au-delà d'une session et ne constitue pas, par elle-même, une biographie[1]. Un LLM contemporain dispose d'un cahier de brouillon plus grand qu'un LLM de 2022 ; cela ne suffit pas à dire qu'il possède une mémoire.
3. Le troisième est celui de l'espace de prédiction. Prédire le prochain token impose un espace cible : celui de la tokenisation. Or beaucoup de ce que l'on cherche à prédire (états physiques, trajectoires biologiques, configurations spatiales, évolution clinique, réponse d'un système sous intervention) ne se réduit pas proprement à une séquence symbolique. La question pertinente n'est plus seulement : quel est le prochain token ? Elle devient : dans quel espace faut-il prédire pour apprendre les invariants utiles ?

Trois familles d'architectures répondent, chacune partiellement, à ces manques :

1. Les world models attaquent la dynamique.
2. Les modèles à mémoire attaquent la persistance.
3. Les architectures de type JEPA attaquent l'espace de prédiction.

Cette correspondance terme à terme est, on le verra, une posture initiale plutôt qu'un état d'équilibre : les frontières se déforment dès que l'on regarde les architectures les plus récentes.

Ces lignées sont souvent présentées comme des alternatives concurrentes aux LLM ou les unes aux autres. Cette présentation est commode pour le marketing, donc naturellement intellectuellement suspecte. Elle est surtout architecturalement fausse. Les trois familles ne se substituent pas proprement. Elles répondent à des déficits différents. Leur trajectoire probable n'est donc pas l'élimination mutuelle, mais la composition de ces architectures. Cette composition, toutefois, ne résout pas automatiquement le problème : elle le déplace vers la coordination des modules, la stabilité d'entraînement, la gouvernance des actions et la validation des sorties. Une architecture hybride n'est pas une synthèse magique ; c'est une pile de problèmes mieux localisés.

Cette note ne traite pas des produits ni du marché. Elle ne tranche pas non plus la question, encore spéculative, de savoir si l'une de ces lignées constitue une voie vers une intelligence générale. Elle propose une cartographie technique minimale pour éviter trois confusions : appeler mémoire un long contexte, appeler world model tout système qui semble comprendre le monde, et appeler JEPA une alternative générale aux LLM alors qu'il s'agit d'abord d'une autre cible de prédiction.

2. Définitions opératoires et coupures

Une définition utile n'est pas une définition séduisante. C'est une définition qui permet de trancher, de se positionner. Les trois termes qui suivent sont utilisés de manière trop large dans le débat public. Il faut donc les restreindre sans prétendre abolir tous leurs autres usages.

Un world model, au sens strict, est un modèle qui apprend la dynamique d'un environnement à partir d'observations, généralement couplées à des actions, et qui permet de prédire un état futur conditionnellement à une séquence d'actions. Cette définition stricte met l'accent sur la projection dynamique : que devient l'environnement si l'agent fait ceci plutôt que cela ? Elle s'applique clairement à la lignée Ha et Schmidhuber, puis Dreamer, où un agent apprend à planifier dans un espace latent plutôt que par essais répétés dans l'environnement réel.

Il existe toutefois un usage plus large du terme. Des modèles vidéo génératifs comme Sora, ou des systèmes interactifs comme Genie, peuvent être décrits comme des world models implicites lorsqu'ils apprennent des régularités temporelles, physiques ou spatiales sans disposer nécessairement d'actions explicites annotées. Dans ce cas, la dynamique est apprise, mais l'action peut être latente, induite ou reconstruite. La distinction importe : un world model strict est conditionné à l'action ; un world model implicite apprend une dynamique du monde sans que cette conditionnalité soit toujours explicite. Confondre les deux permet de faire de belles annonces et de mauvais choix d'architecture, ce qui est une tradition industrielle désormais bien établie.

Un modèle à mémoire est une architecture qui distingue le calcul courant d'un stockage persistant, compressé ou réadressable. Le critère n'est pas la longueur du contexte. Le critère est la différenciation entre traitement immédiat et conservation. Une base externe indexée, un état récurrent compressé et un module de mémoire appris sont trois mécanismes très différents ; ils partagent seulement l'idée qu'une partie de l'information doit survivre à l'inférence immédiate ou être réutilisée dans une séquence longue.

JEPA, pour Joint Embedding Predictive Architecture, désigne une famille d'architectures qui prédisent dans l'espace des représentations plutôt que dans l'espace brut des observations. Une vue contexte et une vue cible sont encodées ; un prédicteur apprend à approcher la représentation de la cible à partir de la représentation du contexte. La perte est calculée dans l'espace latent, non sur une reconstruction pixel par pixel. La proposition centrale est donc simple : pour apprendre ce qui compte, il faut éviter de forcer le modèle à reconstruire ce qui est visible mais non pertinent.

Ces définitions produisent trois coupures.

1. Première coupure : prédire des observations ou prédire des représentations. Elle sépare les world models génératifs, qui peuvent reconstruire des pixels, des architectures JEPA, qui apprennent des représentations prédictives sans reconstruction explicite de l'observation brute.
2. Deuxième coupure : contexte ou mémoire. Elle sépare les LLM à long contexte des modèles à mémoire au sens architectural. Un contexte long permet d'accéder à davantage d'informations pendant une inférence ; une mémoire impose une structure de conservation, d'écriture, de rappel, d'oubli ou de compression.
3. Troisième coupure : prédiction passive ou prédiction conditionnée à l'action. Elle sépare la continuation linguistique d'un modèle dynamique. Un LLM peut produire une phrase sur les conséquences d'une action ; un world model vise à simuler l'effet de cette action sur un état latent ou observable.

Ces coupures ne sont pas des frontières absolues. Elles sont des instruments d'analyse. Leur fonction n'est pas de figer le paysage, mais d'empêcher la confusion des niveaux.

3. Les world models génératifs

La famille des world models génératifs est la plus ancienne des trois lignées discutées ici. Elle est aussi la plus directement liée à la planification et au contrôle.

Ha et Schmidhuber publient en 2018 un article explicitement intitulé World Models. L'architecture y est simple : un module perceptuel, souvent un autoencodeur variationnel[S3], compresse l'observation en un vecteur latent ; un modèle dynamique récurrent apprend à prédire les états latents futurs ; un contrôleur choisit les actions à partir de cet état latent. L'idée importante n'est pas seulement la compression. Elle est que le contrôleur peut être entraîné dans le « rêve » du modèle, c'est-à-dire dans la simulation interne apprise par le système.

La lignée Dreamer généralise cette intuition. DreamerV1 introduit un Recurrent State Space Model combinant état déterministe et état stochastique, afin de modéliser à la fois continuité temporelle et incertitude. DreamerV2 démontre que la planification dans un espace latent[S4] peut rivaliser avec des méthodes de renforcement plus directement ancrées dans l'environnement. DreamerV3 renforce la stabilité et l'étendue d'application du paradigme. DayDreamer transpose cette logique vers des robots physiques, où l'apprentissage par imagination permet de réduire le coût et le risque de l'essai-erreur dans le monde réel.

À une autre échelle, les modèles vidéo génératifs récents peuvent être lus comme des world models implicites. Lorsqu'un système apprend à produire des séquences vidéo cohérentes, il doit internaliser certaines régularités : persistance des objets, occlusions, continuité spatiale, gravité apparente, trajectoires plausibles. Cela ne signifie pas qu'il

possède une physique explicite du monde. Cela signifie qu'une partie de la dynamique visible est capturée dans ses représentations. L'ambiguïté commence exactement ici : entre régularité apprise et modèle d'intervention, entre cohérence visuelle et simulation contrôlable.

Le mécanisme commun de la famille stricte est néanmoins clair. L'agent ne planifie pas directement dans l'environnement ; il planifie dans une approximation interne de l'environnement. La boucle de contrôle évalue des trajectoires futures dans un espace latent[S4], puis sélectionne une action en fonction de ces trajectoires imaginées. C'est le déplacement fondamental : apprendre moins dans le monde réel, apprendre davantage dans le modèle du monde.

Les limites sont structurelles.

1. La première est computationnelle. Lorsque la supervision passe par la reconstruction d'observations riches, une partie considérable de la capacité est dépensée à modéliser des détails peu pertinents pour la décision : textures, bruit, micro-variations visuelles, éléments contingents. La reconstruction pixel peut devenir un mauvais professeur : très exigeante, peu sélective, dépensant la capacité du modèle sur des détails sans portée décisionnelle.
2. La deuxième est temporelle. Les erreurs de prédiction se composent. À court horizon, un rollout latent peut être utile ; à long horizon, l'accumulation d'erreurs déforme progressivement la trajectoire. Ce problème n'est pas un détail d'optimisation. Il affecte la fiabilité même de la planification.
3. La troisième est distributionnelle. Un world model apprend une dynamique locale à la distribution d'entraînement. Si l'environnement réel diverge structurellement, la simulation interne devient trompeuse. C'est le problème classique du transfert sim-to-real, mais il prend ici une forme plus générale : tout modèle dynamique est fiable dans un domaine de validité, pas dans le monde en soi.

Ces limites n'invalident pas les world models. Elles définissent leur domaine d'emploi. Elles expliquent aussi pourquoi la question de l'espace de prédiction devient centrale.

4. JEPA et la rupture représentationnelle

JEPA répond à une faiblesse précise des modèles génératifs : l'obligation de prédire dans l'espace brut des observations. Sa proposition n'est pas de reconstruire mieux. Elle est de ne pas reconstruire ce qui ne mérite pas de l'être.

Dans le texte de positionnement de Yann LeCun, l'argument est direct. Une grande partie du monde observable est imprévisible dans ses détails fins et inutile pour l'action. La position exacte de chaque feuille dans un arbre, la texture d'un mur, le bruit de fond d'une image ou la micro-variation d'un éclairage ne constituent pas nécessairement des variables pertinentes pour apprendre la structure du monde. Un modèle qui consacre

une partie importante de sa capacité à reconstruire ces détails gaspille une ressource qui pourrait être utilisée pour apprendre des invariants.

JEPA déplace donc la cible. Il ne prédit pas l'observation ; il prédit la représentation de l'observation. Une partie de l'entrée sert de contexte. Une autre partie sert de cible. Deux encodeurs produisent des représentations. Un prédicteur apprend à transformer la représentation du contexte pour approcher celle de la cible. La perte est calculée dans l'espace de représentation. Aucune reconstruction pixel n'est requise.

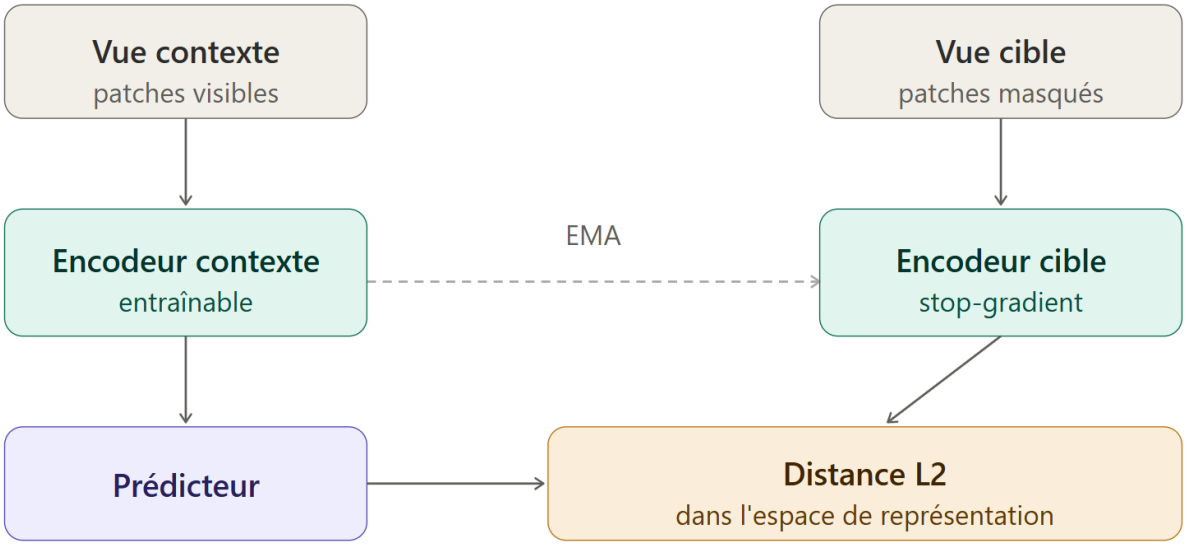

Figure 1. Architecture JEPA. Le contexte (patches visibles d'une image ou d'une vidéo) et la cible (patches masqués) sont traités par deux encodeurs distincts. L'encodeur cible n'est pas entraîné par rétropropagation : ses poids suivent ceux de l'encodeur contexte par moyenne mobile exponentielle[4]. Un prédicteur transforme la représentation du contexte pour qu'elle approche celle de la cible, et la perte est calculée (c'est le point théorique central) dans l'espace de représentation, jamais dans l'espace des pixels[S5].

Ce déplacement est plus profond qu'il n'en a l'air. Le modèle n'apprend plus à produire une image plausible. Il apprend à produire une représentation compatible avec ce qui devrait être présent. L'objectif n'est donc pas la fidélité sensorielle mais la prédictibilité abstraite. ***C'est pourquoi JEPA peut être compris (de façon « approximative ») comme une tentative d'apprendre les contraintes du monde plutôt que ses apparences***.

Le mécanisme d'entraînement doit toutefois éviter l'effondrement trivial. Si les deux encodeurs apprenaient à produire une constante, la perte pourrait être minimisée sans information utile. Pour éviter cela, l'encodeur cible n'est pas directement mis à jour par rétropropagation ; ses poids suivent ceux de l'encodeur contexte par moyenne mobile exponentielle[S5]. Ce mécanisme de self-distillation stabilise l'apprentissage et empêche le système de résoudre la tâche en supprimant l'information.

I-JEPA applique ce principe aux images. V-JEPA l'étend à la vidéo. V-JEPA-2 ajoute une dimension agentique : le modèle apprend à prédire des représentations futures

conditionnellement à des actions. Ce dernier point change le statut de JEPA. Tant qu'il prédit des représentations masquées, il reste principalement une architecture de représentation prédictive. ***Dès qu'il prédit des états futurs sous condition d'action, il devient un world model au sens strict, mais non génératif***.

Cette évolution est importante car elle montre que JEPA n'est pas une famille séparée une fois pour toutes. C'est une réponse à la question de l'espace de prédiction. Lorsqu'elle est couplée à l'action, elle rejoint la question de la dynamique. La frontière entre JEPA et world model ne disparaît pas ; elle se reformule. ***La distinction pertinente devient : world model génératif dans l'espace des observations, ou world model prédictif dans l'espace des représentations.***

Les limites de JEPA doivent être énoncées sans complaisance.

1. La première concerne l'échelle. Les architectures JEPA n'ont pas démontré, à ce jour, une loi d'échelle comparable à celle observée pour les grands modèles de langage. La difficulté n'est pas seulement empirique. Elle est aussi formelle. Le next-token prediction fournit au texte une tâche universelle, massive, homogène et naturellement disponible. JEPA ne dispose pas, à ce stade, d'un objectif aussi universel, homogène et industriellement exploitable : le choix des vues, du masquage, de l'espace latent, des horizons temporels et des pertes reste dépendant du domaine.
2. La deuxième concerne l'absence d'un mécanisme équivalent à l'in-context learning. Les LLM ne se contentent pas de stocker des connaissances dans leurs paramètres ; ils apprennent à utiliser le contexte comme programme local. JEPA apprend des représentations prédictives, mais il n'a pas encore démontré une capacité générale comparable à reconfigurer son comportement à partir d'une séquence d'exemples arbitraires.
3. La troisième concerne la portée modale. JEPA est particulièrement naturel pour la perception (« visuelle ») : image, vidéo, éventuellement robotique. Il ne remplace pas un modèle linguistique général. Le présenter comme une alternative globale aux LLM est une facilité rhétorique. JEPA traite un autre angle mort : la prédiction représentationnelle. Il ne résout pas, à lui seul, la mémoire, le raisonnement symbolique, la gouvernance de l'action ou la génération linguistique.

JEPA est donc une proposition forte, mais partielle. Son intérêt ne tient pas à la promesse de remplacer les LLM. Il tient à la possibilité de sortir l'apprentissage prédictif de l'espace étroit du token et de la reconstruction pixel.

5. Les modèles à mémoire

La mémoire est probablement le terme le plus maltraité du débat contemporain sur l'IA. Le contexte long est appelé mémoire. Une base documentaire est appelée mémoire. Un état caché est appelé mémoire. Une préférence utilisateur persistée est appelée mémoire. À ce stade, le mot couvre tellement d'objets différents qu'il a cessé de désigner quoi que ce soit de précis.

Une taxonomie minimale distingue trois sous-familles.

1. La première est la mémoire externe à index. Le modèle principal reste généralement un LLM auto-régressif. Un système externe récupère des documents, passages, fragments ou événements pertinents, puis les injecte dans le contexte. C'est le principe du RAG. MemGPT ajoute une couche de gestion plus explicite : le système décide quoi sauvegarder, quoi rappeler, quoi résumer. Cette sous-famille a un avantage majeur : elle est gouvernable. Les contenus peuvent être inspectés, supprimés, versionnés, tracés, soumis à des règles d'accès. Sa limite est symétrique : elle suppose que la mémoire utile peut être convertie en objets indexables, souvent textuels. Un état physiologique continu, une trajectoire clinique probabiliste ou une dynamique de système s'y logent mal.
2. La deuxième est la mémoire d'état récurrente compressée. Mamba, RWKV et plus largement les state space models[S2] maintiennent un état caché de taille fixe mis à jour au fil de la séquence. La mémoire n'est pas externe. Elle est dans l'état. L'avantage est computationnel : le coût peut croître linéairement avec la longueur, là où l'attention transformer classique devient coûteuse. La limite est informationnelle : compresser, c'est choisir ; choisir, c'est oublier. Ce qui n'est pas retenu dans l'état courant ne peut pas être récupéré plus tard par simple retour au contexte. La mémoire est continue mais « lossy ».
3. La troisième est la mémoire long-terme apprise. Titans illustre cette direction : un module de mémoire neuronal apprend quoi écrire, quand écrire, comment oublier et comment réutiliser. L'architecture distingue mémoire de travail, mémoire long-terme et parfois mémoire persistante. La mémoire cesse d'être seulement un stockage externe ou un état implicite ; elle devient un composant entraîné.

Cette tripartition peut être rapprochée de catégories issues de la psychologie cognitive (mémoire de travail, mémoire de référence, mémoire épisodique), mais l'analogie doit rester strictement limitée. Les architectures actuelles n'implémentent pas une mémoire humaine. Elles implémentent des mécanismes de conservation, récupération ou compression de l'information.

Une mémoire épisodique, au sens architectural minimal, suppose trois conditions : Un événement singulier indexé temporellement, un rappel orienté par la situation présente, et une mise à jour contextuelle qui ne soit ni simple surécriture ni écrasement.

Aucune des trois sous-familles industrielles courantes ne satisfait pleinement ces trois conditions. Certains modules s'en approchent ; aucun ne mérite qu'on confonde stockage persistant et mémoire au sens propre.

Le point architectural est donc précis. Les modèles à mémoire ne résolvent pas le problème du monde. Ils résolvent une partie du problème de la persistance. Ils permettent de conserver ou de compresser des traces, pas nécessairement de les comprendre, de les hiérarchiser ou de les utiliser causalement.

6. Synthèse comparative

Les trois familles peuvent maintenant être lues en regard. La matrice ci-dessous formalise cette mise en regard sur six axes architecturaux. Elle ne constitue pas un guide de décision : elle est un outil de cartographie, dont la fonction est d'éviter la confusion conceptuelle, non d'arbitrer un choix industriel.

Critère	LLM référence	World models génératifs	JEPA	Modèles à mémoire
Objet prédit	Token suivant	Pixel / observation	Représentation	Token suivant
Espace cible	Symbolique	Latent + reconstruction	Latent uniquement	Symbolique
Action	Non	Centrale	V-JEPA-2 uniquement	Non
Mémoire	Contexte plat	État récurrent	Aucune dédiée	Index, état, apprise
Évaluation	Perplexité, benchmarks	Reconstruction, return RL	Linear probing	Recall, retrieval
Coût	Quadratique en longueur	Reconstruction pixel	Self-distillation	Selon sous-famille

Figure 2. Cartographie comparative. Quatre familles, six critères. La colonne grisée est posée comme référence, non comme membre de la taxonomie.

La lecture habituelle d'une telle matrice est positive : on demande, pour chaque famille, ce qu'elle sait faire. La lecture utile est inverse. Ce qui éclaire le débat n'est pas la liste des capacités revendiquées, mais celle des capacités structurellement absentes, non par défaut d'ingénierie mais par construction architecturale.

Le LLM auto-régressif ne peut pas apprendre la dynamique d'un environnement par sa seule tâche d'entraînement. Aucun volume de texte, aussi grand soit-il, n'expose le modèle à la conditionnalité d'une intervention. Cette absence n'est pas une lacune temporaire que des paramètres supplémentaires combleraient ; elle est consubstantielle à l'objectif d'apprentissage. Le texte décrit des actions et leurs conséquences, mais il ne fait pas agir le modèle dans un environnement où ses propres choix modifieraient les observations suivantes.

Le world model génératif ne peut pas, à lui seul, exploiter sa simulation interne pour des séquences arbitrairement longues. La composition d'erreurs sur l'horizon est une propriété mathématique du chaînage prédictif, non un défaut de calibration. Tout horizon utile est un horizon borné. Cette borne n'est pas un obstacle à lever mais une caractéristique à respecter dans la conception du système qui appelle le world model, typiquement par re-planification fréquente plutôt que par confiance prolongée dans le rollout.

JEPA ne peut pas, dans son état actuel, prédire dans l'espace symbolique avec la flexibilité du next-token prediction. Sa proposition centrale, prédire dans l'espace des représentations, exclut par construction la production de tokens explicites. Le passage de la représentation à l'action linguistique reste à inventer, et il n'est pas certain qu'il puisse être inventé sans réintroduire un décodeur génératif. Tant que cette articulation n'est pas résolue, JEPA s'adresse principalement à la perception et à la dynamique non textuelle.

Un modèle à mémoire ne garantit pas, par sa seule architecture, que ce qu'il conserve sera pertinent, légitime ou réutilisable dans le bon contexte. Il peut apprendre des politiques de stockage ou de rappel, mais la pertinence reste dépendante de l'objectif, du domaine et du régime de gouvernance.

Cette lecture inversée fait apparaître ce que la matrice ne peut pas dire : aucune des quatre familles n'est défaillante au sens où elle échouerait à sa tâche. Chacune fait ce qu'elle a été conçue pour faire. La question architecturale n'est donc pas de mesurer leur performance individuelle ; elle est de comprendre ce qu'elles refusent par nature, pour pouvoir composer ce qu'aucune n'apporte seule.

Trois axes échappent par ailleurs à la matrice et doivent être réintroduits explicitement avant tout usage industriel. Le premier est la maturité de l'écosystème, profondément asymétrique :

- LLM industrialisés,
- world models en R&D avancée principalement académique,
- JEPA en preuve de concept représentationnelle,
- modèles à mémoire en industrialisation hétérogène par sous-famille.

Le deuxième est le coût d'intégration multi-modules, qui croît plus vite que le nombre de modules, chaque interface étant elle-même un objet de validation.

Le troisième est l'ensemble des contraintes de déploiement, gouvernance, auditabilité, conformité, qui appartiennent au contexte d'usage et non à l'architecture nue. Une matrice élégante peut conduire à un système ingérable.

La question « laquelle va gagner ? » est donc mal posée. Elle suppose une concurrence globale là où il existe des fonctions différentes. Elle transforme une décision d'architecture en pari tribal, ce qui est une méthode de gouvernance certes populaire, mais rarement productive. La lecture correcte se fait par déficit adressé : au langage et à la manipulation symbolique répond le LLM ; à la projection dynamique sous action répond le world model ; à la prédiction représentationnelle sobre répond JEPA ; à la persistance, au rappel et à la compression temporelle répond le modèle à mémoire. Cette liste ne suffit pas à choisir une solution ; elle suffit à ne plus se tromper sur le problème.

7. Convergences en cours

Les trois lignées ne restent pas séparées. Elles convergent. Cette convergence est réelle, mais elle ne doit pas être confondue avec une synthèse déjà accomplie.

1. Premier mouvement : JEPA devient agentique. Avec V-JEPA-2, la prédiction représentationnelle n'est plus seulement liée au masquage perceptif. Elle devient conditionnée à des actions. JEPA entre alors dans le territoire des world models stricts, mais par un chemin non génératif : il ne simule pas nécessairement des pixels futurs ; il prédit des représentations futures utiles.
2. Deuxième mouvement : les transformers deviennent mémoriels. Titans, MemGPT et diverses architectures hybrides signalent une même pression : le contexte ne suffit pas. Il faut distinguer ce qui est manipulé maintenant, ce qui doit être conservé, ce qui doit être rappelé, ce qui doit être oublié. La mémoire devient un composant d'architecture, non une simple augmentation de longueur de séquence.
3. Troisième mouvement : les world models intègrent des mémoires et des représentations plus abstraites. Dreamer possède déjà un état latent récurrent pouvant être lu comme mémoire de travail. La question ouverte est celle du couplage entre un modèle dynamique latent, une mémoire long-terme apprise et un encodeur de représentations prédictives. C'est probablement l'une des directions les plus importantes pour les architectures orientées action.

Mais cette convergence déplace la difficulté. Composer un LLM, un world model, une mémoire et un encodeur prédictif ne produit pas automatiquement un système

supérieur. ***Cela produit un système plus difficile à entraîner, à interpréter, à valider et à gouverner***.

Quatre problèmes apparaissent immédiatement.

1. Le premier est la propagation d'erreurs. Une représentation approximative alimente une mémoire partielle, qui alimente une planification incertaine, qui produit une action dont les conséquences retournent dans le système. Dans une architecture composite, l'erreur ne reste pas locale. Elle circule.
2. Le deuxième est la coordination des objectifs. Un module peut optimiser la fidélité représentationnelle, un autre la performance de planification, un troisième la pertinence du rappel, un quatrième la génération linguistique. Rien ne garantit que ces objectifs soient alignés, et en pratique ils ne le sont pas spontanément.
3. Le troisième est la validation. Une architecture modulaire exige de valider les composants, leurs interfaces et leurs interactions. La surface de test augmente plus vite que le nombre de modules. C'est l'un des nombreux endroits où l'enthousiasme architectural meurt, poignardé par l'assurance qualité.
4. Le quatrième est la gouvernance. Qui décide qu'une représentation est suffisamment fiable pour alimenter une action ? Quelle mémoire peut être mobilisée ? Quel événement doit être loggé ? Quelle action doit être refusée ? À partir de quel seuil l'humain doit-il reprendre la main ? Ces questions ne sont pas périphériques. Elles deviennent centrales dès qu'une architecture composite agit dans un environnement réel.

La convergence est donc bien la direction probable. Mais elle ne doit pas être racontée comme la résolution du problème. Elle est l'ouverture d'un problème de niveau supérieur.

Dans des terrains comme les jumeaux numériques cliniques, cette conclusion n'est pas théorique. Un système qui reconstruit ou projette des trajectoires patient doit combiner un modèle de dynamique, une mémoire de l'historique clinique, une représentation abstraite des états non observés et une couche de gouvernance. Aucune des trois familles ne suffit seule. Mais leur composition n'est acceptable que si les interfaces, les hypothèses et les limites de validité sont explicites ; sans quoi le système hérite des angles morts de chaque module sans hériter de leurs garanties.

8. Limites authentiques

Cette cartographie a elle-même des limites. Les expliciter est nécessaire, non par prudence décorative, mais parce qu'une carte qui ne dit pas ce qu'elle exclut devient vite une idéologie.

1. La première limite est l'absence de benchmark inter-familles. Les LLM sont évalués par perplexité, benchmarks linguistiques, raisonnement symbolique ou tâches de programmation. Les world models sont évalués par rendement en

renforcement, erreur de prédiction ou performance de contrôle. JEPA est évalué par linear probing, k-NN, transfert de représentations ou performances aval. Les systèmes à mémoire sont évalués par rappel, exactitude contextuelle ou capacité à exploiter des séquences longues. Ces protocoles ne mesurent pas la même chose. Les comparaisons directes sont donc rarement scientifiques. Elles sont souvent éditoriales, y compris, à plus petite échelle, celle de cette note.

2. La deuxième limite est l'absence de loi d'échelle démontrée pour JEPA au niveau des grands LLM. Ce point ne disqualifie pas l'approche, mais il interdit d'en faire une alternative déjà prouvée. La différence entre une direction prometteuse et un paradigme dominant s'appelle la preuve quantitative. C'est pénible, mais la réalité a parfois ce mauvais goût.
3. La troisième limite est la fragilité du transfert sim-to-real pour les world models. Même lorsque l'apprentissage en simulation est performant, le passage vers un environnement physique, clinique ou organisationnel réel introduit des écarts de distribution, des variables non observées et des contraintes d'action qui ne se laissent pas absorber par une simple augmentation de données.
4. La quatrième limite est l'ambiguïté du terme modèle du monde. Il désigne tantôt un modèle dynamique conditionné à l'action, tantôt une représentation perceptive riche, tantôt un système génératif impressionnant, tantôt une métaphore cognitive. Cette polysémie est utile pour vendre une vision ; elle est dangereuse pour concevoir une architecture.
5. La cinquième limite est plus rarement formulée : aucune de ces familles n'intègre nativement une gouvernance complète. Un LLM ne sait pas naturellement distinguer recommandation et action. Un world model ne sait pas naturellement borner son domaine de validité. JEPA ne fournit pas par lui-même une traçabilité causale de ses représentations. Un module de mémoire ne garantit pas que ce qu'il rappelle est légitime, actuel ou autorisé. La gouvernance doit donc être architecturée autour du modèle, et parfois dans le modèle, mais elle n'émerge pas automatiquement de la performance.

Ce point est décisif en environnement régulé. Un système composite qui simule, rappelle, prédit et agit doit exposer ses hypothèses, ses événements, ses refus, ses seuils, ses incertitudes et ses responsabilités. À défaut, on obtient une architecture techniquement séduisante et réglementairement inutilisable.

9. Conclusion

Le débat sur les architectures post-LLM est souvent formulé comme une succession de remplacements. Les LLM auraient remplacé les modèles symboliques. Les world models remplaceraient les LLM. JEPA remplacerait les modèles génératifs. Les modèles à mémoire répareraient la faiblesse du contexte. Cette lecture est trop simple.

Le bon diagnostic est fonctionnel. Le paradigme auto-régressif a révélé une puissance remarquable dans la manipulation des séquences symboliques, mais il laisse ouverts trois problèmes : la dynamique du monde, la persistance des informations et le choix de l'espace de prédiction. Les world models, les modèles à mémoire et JEPA répondent à ces trois problèmes, chacun partiellement, chacun avec ses limites.

Un LLM parle du monde. Un world model projette des états possibles du monde. JEPA apprend des représentations prédictives du monde. Un modèle à mémoire conserve ou rappelle des traces du monde. Aucun ne constitue, seul, une architecture complète.

La question stratégique n'est donc pas : quel paradigme va gagner ? Elle est : quelle combinaison minimale de capacités est nécessaire pour le cas d'usage considéré, sous quelles hypothèses, avec quel domaine de validité, quel coût, quelle gouvernance et quelle preuve ?

Cette reformulation change la nature de la décision. Elle interdit de confondre annonce de laboratoire, promesse produit et architecture exploitable. Elle impose de raisonner par fonction, par interface et par validation. C'est moins spectaculaire qu'une prophétie. C'est surtout le seul niveau où une décision d'architecture cesse d'être une croyance et devient défendable.

Footnotes

1. Sur la distinction entre mémoire biographique et stockage informationnel, voir J. Vetillard, *Encodage, transduction et modèles du monde*, Twingital Institute, 2025.
2. *State Space Models*. Famille d'architectures qui maintiennent un état caché de taille fixe mis à jour à chaque pas de temps. À la différence du transformer classique, leur coût peut croître linéairement avec la longueur de séquence.
3. *Variational Autoencoder*. Réseau qui apprend à compresser une observation, par exemple une image, en un vecteur de dimension réduite appartenant à l’espace latent, tout en préservant l'information utile à la reconstruction.
4. L’espace latent est un espace auxiliaire de représentation, appris pour rendre certaines opérations (reconstruction, prédiction, contrôle) plus simples. Ce n’est ni un sous-espace du réel, ni une copie miniature du monde. C’est une coordination interne utile, avec une géométrie largement arbitraire et seulement partiellement alignée avec la structure des données.
5. Technique dite *EMA*, ou moyenne mobile exponentielle, par laquelle les poids de l'encodeur cible suivent ceux de l'encodeur contexte avec retard, afin de stabiliser l'entraînement et d'éviter l'effondrement trivial.