Strasbourg, 25 février 2026 T-(2025)3rev2

**COMITÉ CONSULTATIF DE LA CONVENTION POUR LA PROTECTION DES INDIVIDUS EN MATIÈRE DE TRAITEMENT AUTOMATIQUE DES DONNÉES PERSONNELLES**

**(CONVENTION 108)**

# Projet de lignes directrice sur la vie privée et la protection des données dans le contexte des systèmes basés sur les LLM

www.coe.int/dataprotection

# Introduction

# 1. Objectif et portée

1.1 Contexte et Finalité

- Aider les parties prenantes à identifier, évaluer, atténuer et surveiller (approche IAMM) les risques liés aux systèmes basés sur les LLM au titre des articles et principes de la Convention 108+.

1.2 L'approche IAMM et ses étapes

1.3 Public cible et parties prenantes à travers l'écosystème des LLM

- États et governments
- Fournisseurs LLM (développeurs de modèles et concepteurs systèmes)
- Responsable de traitement (intégrateurs de systèmes et fournisseurs de services)
- Industrie
- Organisations de la société civile
- Utilisateurs (c'est-à-dire utilisateurs finaux)
- Régulateurs / Autorités de controle

# 2. Concepts clés et leurs définitions

2.1 Notions clés

2.1.1 Une approche du cycle de vie enracinée dans le Traité sur l'IA et dans l'écosystème LLM

2.1.2 Différence entre le modèle LLM et le système basé sur LLM

2.1.3. Cinq étapes fondamentales des processus dynamiques de gestion du cycle de vie et des risques des LLM

2.2 Types de risques pour la vie privée tout au long du cycle de vie de l'IA et dans l'écosystème LLM

2.2.1 Cartographie des risques de confidentialité dans le contexte des systèmes basés sur les LLM

2.2.2 Risques de vie privée dans l'écosystème LLM des données

2.3 Risques émergents pour la vie privée

## 3. Convention 108+ Principes et articles pertinents aux systèmes basés sur LLM

3.1 Comprendre les principes de la Convention dans le contexte des dernières évolutions technologiques des LLM et des cadres agents

- Concentrez-vous sur les principes fondamentaux de la vie privée issus de la Convention 108+ [cf. chapitres 2 et 3].
- Expliquez et contextualisez comment ces principes se rapportent spécifiquement aux grands modèles de langage et aux systèmes basés sur des LLM ainsi qu'à leur cycle de vie (par exemple, lors de la formation, de l'ajustement fin ou du déploiement).
- Mettez en lumière les tensions et défis spécifiques des LLM et des systèmes basés sur les LLM soulevés sous ces principes (par exemple, la réutilisation de données, les hallucinations, l'inférence de données personnelles, les risques liés aux données synthétiques).

3.1.1 Sécurité, exactitude et qualité des données, questions de transparence et de responsabilité chez les agents IA

3.1.2 Légalité et équité d'utilisation : le cas des proxys de données utilisés pour inférer les données personnelles et reconstruire la vie privée des individus

3.1.3 Minimisation des données et droits des sujets des données dans l'économie de l'intention dans l'économie de l'intention

3.1.4 Limitation de but dans les paramètres d'agrégation de données multimodales

3.2 Comprendre les articles de la Convention dans le contexte des dernières évolutions technologiques des LLM et des cadres agents

- Structure bipartite citant les articles pertinents de la Convention et incluant dans chaque sous-partie (a) l'interprétation et l'explication de l'article, et (b) sa contextualisation technologique à travers des exemples.

3.2.1 Article 5 : Méthodes d'extraction des données personnelles et mémorisation
3.2.2 Article 6 : Traitement des catégories spéciales de données
3.2.3 Article 7 & Sécurité des données : lorsque les LLM divulguent des données personnelles
3.2.4 Article 8 & Transparence du traitement

3.2.5 Article 9 & Droits des personnes
concernées 3.2.6 Article 10 & minimisation du risque d'ingérence dans les droits et libertés fondamentales des personnes concernées
3.2.7 Article 14 : Flux transfrontaliers de données personnelles.

## 4. Directives spécifiques aux parties prenantes

Pour chaque groupe de parties prenantes (fournisseurs, déployateurs, utilisateurs, régulateurs), fournir :

4.1 Transformation des principes de la Convention 108+ en actions

- Montrez comment chaque principe se traduit par des obligations ou responsabilités concrètes envers ce groupe

4.2. Responsabilités en gestion des risques

- Indiquez comment chaque acteur est censé :
  - Identifier les risques pour la vie privée
  - Évaluer l'impact
  - Appliquer des stratégies d'atténuation
  - Surveillez et évaluez l'efficacité

4.3. Stratégies d'atténuation et bonnes pratiques (par stade de risque et de cycle de vie)

- *Exemple :*
  - *Fournisseurs : pratiques de curation des données, confidentialité différentielle, indicateurs d'évaluation robustes*
  - *Responsable de traitement : utilisation de mécanismes de consentement, surveillance rapide et contrôles d'accès*
  - *Utilisateurs : conception responsable des prompts, cas d'utilisation respectueux de la vie privée*
  - *Régulateurs : mécanismes de surveillance, renforcement des capacités et critères d'audit*

## 5. Considérations de mise en œuvre

5.1 Mettre en avant les mécanismes de gouvernance et de responsabilité
5.2 Promouvoir la collaboration interfonctionnelle (équipes techniques, juridiques, éthiques)
5.3 Encourager les évaluations d'impact sur les droits humains et les droits fondamentaux
5.4 Pointer vers l'interopérabilité avec d'autres cadres réglementaires (par exemple, loi sur l'IA, RGPD, DSA)

## 6. Annexes

Annexe I : Processus de gestion des risques pour les LLM

- Aperçu de l'identification, analyse, atténuation et suivi des risques
- Intégration avec les évaluations d'impact sur la protection des données (DPIA)

Annexe II : Phases du cycle de vie des systèmes LLM

- Description détaillée des étapes du cycle de vie pertinentes pour les LLM

Annexe III (optionnelle) : Études de cas ou exemples

- Exemples illustratifs des risques de vie privée dans les déploiements réels de LLM
- Comment différents acteurs les ont abordés
- Implémentation de l'IA agente

Annexe IV (facultative) : Glossaire

- Liste des termes clés avec des définitions brèves pour une référence facile

# Introduction

Les technologies émergentes créent des opportunités significatives pour les individus, les organisations et les autorités publiques. En même temps, elles peuvent générer de sérieux risques pour la jouissance effective des droits de l'homme, le fonctionnement de la démocratie et le respect de l'État de droit. Ces dernières années, le rythme du développement technologique a de plus en plus transformé les conditions dans lesquelles la vie privée est exercée et les données personnelles sont traitées, nécessitant une attention renouvelée aux garanties qui protègent les individus dans l'environnement numérique. Dans ce paysage en évolution, les grands modèles de langage (LLM) représentent un développement particulièrement influent, combinant des capacités avancées de traitement du langage avec un déploiement automatisé à grande échelle dans tous les secteurs, amplifiant à la fois les opportunités d'innovation et les défis liés à la protection des droits.
Les LLM constituent une catégorie d'intelligence artificielle conçue pour traiter et générer le langage humain à grande échelle. Entraînés sur d'immenses quantités de données et intégrés dans des infrastructures numériques de plus en plus complexes, ils peuvent être utilisés pour automatiser un large éventail de tâches, de la génération et de la synthèse de contenu au support décisionnel et à l'interaction automatisée. [12345] Parce que leur développement et leur exploitation dépendent du traitement à grande échelle des données, qui peut inclure des informations personnelles et sensibles, les LLM soulèvent des considérations importantes en matière de confidentialité et de protection des données. Leurs caractéristiques techniques, notamment l'opacité, les résultats probabilistes et l'adaptation continue, peuvent remettre en cause les garanties traditionnelles et compliquer l'application efficace des principes établis de protection des données.
Les risques peuvent survenir à différents moments du cycle de vie des systèmes basés sur LLM, du développement du modèle au déploiement système et à l'interaction utilisateur. Ces risques peuvent inclure la conservation ou la reproduction involontaire de données personnelles, un manque de transparence concernant le traitement des données, ainsi que des difficultés à garantir une supervision humaine significative et une responsabilité. Lorsqu'ils sont intégrés dans des environnements de prise de décision automatisés ou semi-automatisés, ces systèmes peuvent également affecter la capacité des individus à exercer efficacement leurs droits.

Au-delà des impacts individuels, le déploiement généralisé de systèmes basés sur les LLM peut également avoir des implications sociétales plus larges, notamment sur l'autonomie, la formation de l'identité, la démocratie, le discours public et la confiance dans les environnements numériques.

À mesure que les LLM continuent de s'étendre dans des secteurs tels que le recrutement, l'éducation, la santé et l'administration publique, garantir des protections efficaces en matière de vie privée et de protection des données devient de plus en plus complexe. Ce paysage technologique en évolution renforce l'urgence de réaffirmer et de mettre en œuvre les principes de la Convention 108+ de manière robuste, adaptative et à l'épreuve de l'avenir.

Ce paysage technologique en évolution renforce la nécessité de réaffirmer et de mettre en œuvre les principes de la Convention 108+ de manière à rester robuste, adaptable et capable de faire face aux risques émergents.

Ces lignes directrices visent à soutenir l'application de la Convention pour la protection des individus en matière de traitement automatique des données personnelles (CETS n° 108, « Convention 108 ») et de son protocole d'amendement CETS n° 223 (désigné dans ce document comme « la Convention », « Convention 108+ ») dans le contexte des systèmes basés sur LLM. Ce faisant, ils cherchent à promouvoir le respect du droit à la vie privée et la protection des données personnelles, tels qu'énoncés dans la Convention européenne des droits de l'homme (article 8). Les Lignes directrices s'appuient sur les travaux antérieurs du Comité consultatif de la Convention 108 et des directives connexes du Conseil de l'Europe sur la protection des données et les technologies émergentes.[1]

Ces lignes directrices s'inscrivent également dans l'approche centrée sur l'humain reflétée dans la Convention-cadre du Conseil de l'Europe sur l'intelligence artificielle et les droits de l'homme, la démocratie et l'État de droit (CETS n° 225) (« la Convention-cadre »), qui affirme la nécessité de sauvegarder les droits de l'homme, la démocratie et l'État de droit tout au long du cycle de vie des systèmes d'IA, y compris la vie privée et la protection des données personnelles.

S'appuyant sur les travaux antérieurs du Comité consultatif de la Convention 108 et des directives du Conseil de l'Europe sur la protection des données et l'intelligence artificielle, ces Lignes directrices offrent un cadre cohérent et prospectif pour identifier, évaluer et traiter les risques liés à la vie privée et à la protection des données dans les systèmes basés sur les LLM.

# Section 1 Objectif et portée

Ces lignes directrices traitent des risques liés à la vie privée et à la protection des données liés à l'utilisation des grands modèles de langage (LLM), en particulier en ce qui concerne les droits et principes consacrés dans la Convention 108 et sa version modernisée, la Convention 108+. Ces risques apparaissent à travers différentes phases du cycle de vie des LLM, allant de l'entraînement et de l'inférence du modèle à l'intégration et au déploiement dans des architectures plus larges de systèmes et d'agents, et ont des implications non seulement pour la protection des données, mais aussi pour la vie privée, la dignité et l'autonomie.

Les Lignes directrices présentent une compréhension fondée sur des preuves de la manière dont les systèmes basés sur les LLM et l'IA agente peuvent interférer avec les droits des individus et proposent une méthodologie structurée pour identifier, évaluer, atténuer et surveiller (IAMM) ces risques dans des contextes réels au sein de divers groupes de parties prenantes.

## 1.1 Contexte et objectif

Les LLM transforment rapidement le paysage numérique en permettant de nouvelles formes d'interaction, d'automatisation et de traitement de l'information. Cependant, ces

développements introduisent des risques liés à la vie privée et à la protection des données qui remettent en cause les garanties juridiques et techniques traditionnelles. Cela inclut, sans s'y limiter, la mémorisation et la reproduction involontaires de données personnelles, la susceptibilité à la manipulation lors de l'inférence, et l'érosion plus large de la vie privée par des identités synthétiques, le profilage, l'opacité dans la prise de décision et l'exploitation des données personnelles pour une automatisation orientée action dans des cadres agents.

Les lacunes persistantes de gouvernance incluent des difficultés à identifier les acteurs responsables dans la chaîne de valeur dans les écosystèmes d'IA en couches, à garantir la transparence des résultats probabilistes et à garantir l'exercice efficace des droits des sujets de données.

L'opacité des données d'entraînement du modèle, des processus d'inférence et des configurations des systèmes en aval peut limiter la capacité des individus à savoir si leurs données personnelles ont été utilisées et comment. Cela soulève des questions sur la manière dont les principes fondamentaux de la Convention 108+, y compris l'équité, la transparence, la responsabilité et les recours efficaces, peuvent être respectés dans la pratique.

Informer les sujets de données et identifier des contrôleurs de données responsables au sein d'architectures en couches et distribuées reste un défi de gouvernance persistant. Lorsque les résultats sont probabilistes et que les configurations destinées à l'utilisateur sont opaques, garantir l'exercice effectif des droits à la vie privée, y compris les droits de rectification, d'objection et d'informations pertinentes concernant le traitement automatisé, devient particulièrement complexe.

L'absence de traces de données observables et accessibles ainsi que d'interfaces utilisateur limite encore davantage la capacité des individus à savoir si et comment leurs données personnelles ont été utilisées, sans parler de contester ou de corriger ce traitement. Cela soulève de sérieuses inquiétudes quant à la possibilité que les dispositions fondamentales de la Convention 108+ puissent être maintenues en pratique sans mesures techniques et organisationnelles complémentaires. Compte tenu du déploiement étendu des systèmes basés sur les LLM et les systèmes agents dans divers contextes, il est urgent de préserver et de faire respecter les droits des sujets de données selon les principes de la Convention 108+ en tenant compte de ces réalités technologiques.

L'objectif de ces lignes directrices est de proposer une approche globale des risques liés à la vie privée et à la protection des données en vertu de la Convention 108+ dans le contexte des LLM et des systèmes basés sur les LLM. Ils :

- clarifier l'application des principes de la Convention 108+ tout au long du cycle de vie des LLM ;
- identifier les principaux risques liés à la vie privée et à la protection des données survenant à différents stades du cycle de vie ; (Section 2)
- fournir une méthodologie structurée pour identifier, évaluer, atténuer et surveiller ces risques ;

- traiter des rôles et responsabilités des parties prenantes concernées, y compris les prestataires, les déployeurs, les autorités publiques, les autorités de supervision et les utilisateurs finaux (Section 1) ; et
- promouvoir la cohérence avec les instruments existants du Conseil de l'Europe et les méthodologies d'évaluation des risques.

Les Lignes directrices adoptent une méthodologie basée sur le cycle de vie reflétant la nature dynamique et évolutive des LLM et des systèmes basés sur les LLM, et reconnaissant que les risques liés à la vie privée et à la protection des données peuvent apparaître à différentes étapes du cycle de vie (Section 2).
Compte tenu du rôle du Comité consultatif et de son travail antérieur dans l'interprétation de la Convention 108+ dans le contexte des technologies émergentes, ces Lignes directrices informent les principes juridiques fondamentaux consacrés dans la Convention 108+ et leur application aux LLM et aux systèmes basés sur des LLM tout au long de leur cycle de vie (Section 3).
Les Lignes directrices adoptent une approche spécifique aux parties prenantes lorsqu'elles développent des obligations ou responsabilités concrètes pour des groupes spécifiques (fournisseurs, déployeurs, utilisateurs, régulateurs). Ils introduisent une compréhension fondée sur des preuves de la manière dont les systèmes d'IA basés sur les LLM et les systèmes d'IA agents peuvent interférer avec les droits à la vie privée et à la protection des données des individus, et illustrent les stratégies d'atténuation correspondantes (Section 4).
Dans l'ensemble, les Lignes directrices fournissent un cadre actualisé, structuré, informé par la recherche et les risques pour comprendre et gérer les risques associés aux systèmes basés sur les LLM. Ils posent les bases d'outils de gouvernance qui mettent en œuvre à la fois les principes et les obligations contraignantes de la Convention 108+.
Une caractéristique centrale de l'application opérationnelle de la Convention 108+ dans ce contexte est la méthodologie de gestion des risques basée sur le cycle de vie énoncée à la Section 4. Les Lignes directrices introduisent une approche structurée de la gouvernance des risques de vie privée basée sur quatre étapes interdépendantes : « Identifier », « Évaluer », « Atténuer » et « Surveiller » (« approche IAMM »). Cette méthodologie vise à soutenir les parties prenantes diverses dans la mise en œuvre des obligations de la Convention 108+ dans des contextes réels, garantissant que les technologies émergentes restent alignées sur les droits fondamentaux, les valeurs démocratiques et l'État de droit.
Les Lignes directrices contribuent aux efforts de normalisation du Conseil de l'Europe pour promouvoir une approche intégrée de la protection des données et de la gouvernance de l'IA en promouvant une approche proactive et fondée sur les droits de l'innovation tout en préservant les principes de transparence, de responsabilité et de dignité humaine. Les Lignes directrices fournissent des orientations transversales sur la gouvernance et les mécanismes de responsabilité qui favorisent la collaboration interfonctionnelle et la surveillance continue (Section 5). Ils complètent les instruments d'évaluation existants tels que les évaluations d'impact sur la vie privée (en vertu de l'article 10 de la Convention 108+) et HUDERIA – la méthodologie d'évaluation de l'impact des droits de l'homme, de la démocratie et de l'État de droit du Conseil de l'Europe, en situant les risques liés à la vie privée dans une perspective systémique plus large et en favorisant l'interopérabilité avec d'autres cadres réglementaires (Section 5).

## 1.2 L'approche IAMM et ses étapes

### *Approche IAMM*

Ces lignes directrices visent à aborder la manière dont les organisations peuvent gérer les risques liés à la vie privée tout au long du cycle de vie des systèmes basés sur LLM et des architectures agentiques, de la collecte et du développement des modèles au déploiement, au suivi et à la supervision post-commercialisation. Ils répondent à une incertitude généralisée et à l'incohérence des pratiques actuelles ainsi qu'au besoin pressant d'une approche harmonisée et cohérente à l'échelle internationale de la gestion des risques liés à la vie privée. Ce faisant, les Lignes directrices cherchent à réduire les processus internes fragmentés, à renforcer la formalisation des évaluations spécifiques à la vie privée et à soutenir l'application efficace des obligations de protection des données dans des architectures technologiques complexes et en rapide évolution.
L'approche comprend :

- Établir une taxonomie commune (par exemple, clarifier la notion de données personnelles dans le contexte des LLM) à la Section 2.1.
- Analyse des risques connus et émergents de la vie privée au sein de l'écosystème des LLM, dans la section 2.2.
- Analyse de la manière dont les Principes et Articles consacrés dans la Convention 108+ ainsi que leur mise en œuvre et leurs garanties sont remis en question par les derniers développements technologiques, à la Section 3.
- Analyse des outils de gestion des risques, y compris des méthodologies d'évaluation centrées sur la vie privée et les droits humains telles que la méthodologie Huderia et COBRA, dans les sections 2.2 et 4.
- Engager les parties prenantes pour prioriser les pilotes et valider la faisabilité et la pertinence des stratégies d'atténuation proposées dans la Section 4.
- Et, en ancrant les meilleures pratiques sur les contraintes et exigences réelles des entreprises, des décideurs politiques et des praticiens de l'IA, dans la Section 5.

Le postulat fondamental est que les risques liés à la vie privée dans les systèmes basés sur les LLM ne peuvent pas être correctement traités uniquement par des pratiques organisationnelles ad hoc ou des outils de conformité existants. Au lieu de cela, une méthodologie structurée, basée sur le cycle de vie, est nécessaire pour identifier, évaluer et atténuer les risques liés à la vie privée, tant au niveau du modèle que du système, de manière dynamique, évolutive et progressive.
La section 4 présente un cadre de gestion des risques liés à la vie privée aligné sur les principes de la Convention 108+, adapté aux réalités techniques et organisationnelles de l'IA générative. Son champ comprend l'identification des risques liés à la vie privée à différentes phases de développement et de déploiement, une évaluation à deux vitesses des risques au niveau du modèle et du système composite, ainsi qu'une cartographie initiale de stratégies d'atténuation viables ainsi qu'une approche avancée de surveillance dans un cadre nommé d'après ses quatre composantes principales : « Identifier », « Évaluer », « Atténuer » et « Surveiller » (« IAMM ») :

1. Identifier les risques et recommandations pour la vie privée basés sur la Convention 108+. Cette première étape consiste à déterminer quels risques pour la vie privée apparaissent à

différents stades du développement et du déploiement de l'IA, comment ils apparaissent, et à identifier les mesures appropriées conformément à la Convention 108+, aux garanties consacrées à l'article 8 de la Convention européenne des droits de l'homme, ainsi qu'aux exigences en matière de risques et de protection des données énoncées dans la Convention-cadre sur l'intelligence artificielle du Conseil de l'Europe (Traité sur l'IA). en particulier l'article 11 sur la vie privée et la protection des données personnelles et l'article 16 sur un cadre de gestion des risques et de l'impact.

2. Fonder ces lignes directrices sur une méthodologie d'évaluation des risques à deux niveaux permet une approche d'identification et d'évaluation des risques qui distingue les risques et responsabilités au niveau du modèle et du niveau système pour les différentes parties prenantes tout au long du cycle de vie technologique, en particulier lorsqu'on considère les principes fondamentaux énoncés au Chapitre II – Principes fondamentaux pour la protection des données personnelles de la Convention 108+, qui incluent la légalité, limitation de l'objectif, minimisation des données, équité, transparence, droits des personnes concernées, sécurité des données et protection des données sensibles. L'objectif est d'équiper les contrôleurs de données d'outils pour évaluer les risques découlant de différentes étapes du cycle de vie, allant des données d'entraînement, la mémorisation ou les hallucinations au niveau du modèle, ainsi que les risques liés à l'intégration système, à l'interaction utilisateur et aux composants tiers au niveau du déploiement.
3. Étude de stratégies viables d'atténuation et de surveillance dans des applications réelles. Développer une approche structurée des évaluations d'impact sur la vie privée, adaptée aux applications basées sur les LLM et les applications agentiques, sera essentielle, afin de garantir que les risques liés à la vie privée soient systématiquement analysés et traités tout au long du cycle de vie de l'IA. De plus, l'intégration de technologies améliorant la vie privée (PET) telles que la confidentialité différentielle, l'apprentissage fédéré et les techniques de chiffrement, ainsi que l'apprentissage automatique général comme le fine-tuning, ou des techniques d'identification des données personnelles telles que celles-ci, sera évoquée comme des stratégies d'atténuation viables, en gardant toujours un œil sur l'applicabilité, les limites et les implications de gouvernance de ces outils dans des contextes pratiques.

## 1.3 Public cible et parties prenantes à travers l'écosystème des LLM

### *Parties prenantes de l'écosystème LLM*

Pour cartographier les défis concrets auxquels sont confrontées les organisations développant ou utilisant des LLM, le public cible est constitué d'un large éventail d'acteurs à travers le cycle de vie et l'écosystème des LLM, des start-ups, des grands fournisseurs technologiques, des auditeurs technologiques, des déploieurs de technologies privés et publics, des autorités réglementaires et des institutions de recherche.

Les rôles et responsabilités sont répartis et interdépendants entre plusieurs acteurs et parties prenantes. Les mesures d'atténuation des risques liés à la vie privée et à la protection des données sont prises en compte tout au long du cycle de vie, de la création du modèle à l'interaction avec l'utilisateur. Ce système de gouvernance intègre une dimension multipartite avec des responsabilités juridiques et opérationnelles distinctes :

– *États et gouvernements*

La Convention 108+ établit une approche technologiquement neutre et fondée sur des principes, répondant aux défis résultant des technologies émergentes, dont la force opérationnelle est définie par les États au niveau national. La responsabilité principale d'établir ou de maintenir des cadres juridiques et institutionnels pour protéger le droit à la vie privée et à la protection des données des individus est garantie par les États et les gouvernements. Ces lignes directrices fournissent des orientations générales sur le développement, l'endossement et la mise en œuvre de cadres d'évaluation des risques de vie privée fondés sur le contexte pour les systèmes basés sur les LLM et soutiennent la convergence réglementaire pour éviter la fragmentation entre divers secteurs.

– *Fournisseurs de LLM (développeurs de modèles et concepteurs systèmes)*

Une approche du cycle de vie fondée sur la Convention 108+ et les normes de la Convention-cadre garantit que les principes de la vie privée et de la protection des données sont intégrés tout au long du développement et de l'utilisation des systèmes d'IA, plutôt qu'introduits à des étapes ultérieures comme une réflexion après coup. Les développeurs de modèles et les concepteurs de systèmes devraient intégrer la confidentialité dès la conception et par défaut au stade de développement (création de modèle), où les données d'entraînement sont collectées, prétraitées et les modèles sont construits. Ces lignes directrices facilitent la compréhension par les fournisseurs de LLM des concepts clés de protection des données établis par la Convention 108+, de leurs efforts de conformité juridique et d'évaluation des risques tout au long du cycle de vie des systèmes basés sur des LLM.

– *Déploieurs (intégrateurs de systèmes et fournisseurs de services)*

En tant que parties prenantes opérationnelles clés, les intégrateurs de systèmes et les prestataires de services restent responsables de l'application des systèmes basés sur les LLM dans leurs opérations quotidiennes. Même dans le cas du déploiement de systèmes tiers, ils assument la responsabilité légale et la responsabilité des opérations de traitement des données et détectent les menaces potentielles pour la vie privée après le déploiement. Ces lignes directrices soutiennent les efforts de conformité juridique des déploieurs, assistent à l'évaluation des risques de vie privée avant le déploiement du système, et guident l'adoption efficace des technologies de protection de la vie privée (PET) appropriées.

– *Industrie*

Les acteurs du secteur jouent un rôle important en développant des normes techniques et des meilleures pratiques qui facilitent la convergence entre divers secteurs. Dans ce but, les Lignes directrices aideront à aligner la gouvernance et la gestion des risques de vie privée, ainsi qu'à mettre en œuvre des évaluations basées sur le cycle de vie.

– *Organisations de la société civile*

Les organisations de la société civile contribuent efficacement à la sensibilisation aux risques liés à la vie privée et à la mise en question de la conformité juridique des cycles de vie des systèmes, en répondant à la sensibilisation aux risques liés à la vie privée et aux boucles de rétroaction entre parties prenantes et régulateurs. Les Lignes directrices proposeront une approche d'évaluation des risques liés à la vie privée et les meilleures pratiques pour les stratégies d'atténuation aux organisations de la société civile afin de poursuivre leur mission et leur action de sensibilisation.

– *Utilisateurs (utilisateurs finaux)*

Les avancées technologiques rapides ont engendré des impacts négatifs potentiels sur les droits et libertés de l'homme, notamment les droits à la vie privée et à la protection des données. Ces lignes directrices informent sur les défis liés à la vie privée interférant avec les

droits des personnes concernées et fournissent une méthodologie structurée pour « identifier », « évaluer », « atténuer » et « surveiller » ces risques dans des contextes réels pour une variété de parties prenantes.

– *Régulateurs / Autorités de controle*

Les autorités de contrôle de la protection des données et d'autres autorités compétentes contribuent de manière significative à la contrôle de la légalité des opérations de traitement des données tout au long du cycle de vie des systèmes basés sur les LLM ainsi qu'à la conformité légale des cadres de protection des données. Ces lignes directrices les aideront à fixer des seuils de convergence réglementaire pour la transparence et la responsabilité des systèmes, à exercer efficacement des pouvoirs de supervision concernant les systèmes basés sur des LLM et à adopter ou examiner les méthodologies d'évaluation des risques de vie de vie pour le cycle de vie des LLM. Le document les aidera à faciliter la coordination entre différents acteurs.

# Section 2
# Concepts clés et leurs définitions

Cette section offre un aperçu technique de la manière dont les données personnelles peuvent être traitées dans les systèmes basés sur des LLM, établissant la couche conceptuelle fondamentale du cadre d'évaluation des risques proposé. Il clarifie la terminologie clé et examine les implications en matière de vie privée des développements actuels, y compris les questions liées à l'explicabilité et à la transparence, ainsi que leur pertinence pour la gouvernance de l'IA.

Plutôt que de fournir une analyse exhaustive, la section se concentre sur les caractéristiques techniques essentielles et les mécanismes internes qui influencent les risques liés à la vie privée, notamment la manière dont les données sont absorbées, transformées, organisées et potentiellement mémorisées par ces systèmes. Il aborde également les risques liés aux données textuelles d'entrée et à leur représentation. En clarifiant le fonctionnement interne de ces systèmes, nous visons à montrer qu'une culture technique plus approfondie est une condition préalable au développement de garanties efficaces de la vie privée et de mécanismes de gouvernance ancrés dans un comportement réel du système.

## 2.1 Concepts essentiels

2.1.2 Une approche du cycle de vie ancrée dans le Traité sur l'IA et dans l'écosystème LLM

Une analyse introductive des risques de vie privée du cycle de vie des systèmes basés sur LLM peut identifier quatre étapes principales en fonction des types de traitement des données impliqués dans l'entraînement des modèles de fondation, l'adaptation et l'alignement post-entraînement, l'intégration et l'optimisation des systèmes ; et le déploiement et l'interaction des utilisateurs dans des contextes opérationnels spécifiques.

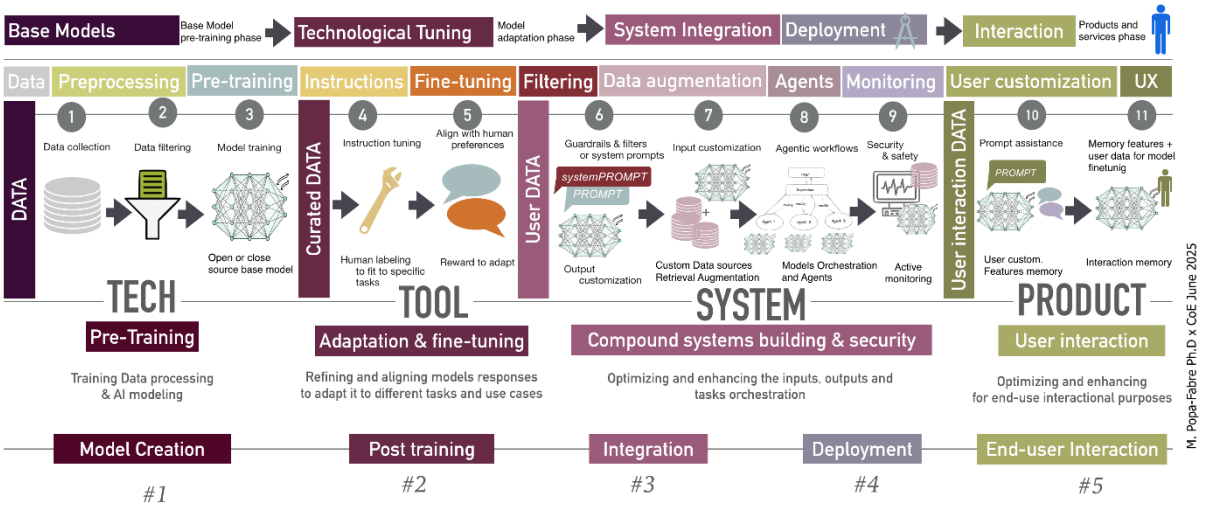


*Figure 1 : Systèmes basés sur LLM sur le cycle de vie/chaîne de valeur, un aperçu large basé sur les pratiques en rapide évolution de la construction de systèmes composés IA pour optimiser les applications basées sur LLM à la phase « Produit ».*

Il est important de noter que, bien que la phase fondamentale du développement des LLM repose généralement sur le traitement de jeux de données très vastes et diversifiés, incluant la collecte et le prétraitement des données (étapes 1 à 3), les étapes suivantes dépendent de plus en plus de données plus soigneusement sélectionnées et spécifiques au contexte. Lors de l'adaptation post-entraînement, les modèles sont affinés et alignés pour des usages

particuliers, souvent par le traitement de jeux de données ciblés et d'entrées spécifiques à chaque tâche. Cette phase intermédiaire, qui comprend l'ajustement fin, l'intégration du système et la préparation au déploiement (étapes 4 à 9), est particulièrement significative du point de vue de la confidentialité. C'est à ce stade que les garanties de protection des données peuvent être intégrées dans la conception des systèmes, les structures de gouvernance et les configurations opérationnelles. Malgré son importance, cette phase est souvent insuffisamment examinée dans les discussions sur la vie privée.

Aux étapes ultérieures du cycle de vie, les pratiques d'optimisation et de personnalisation façonnent davantage le fonctionnement des systèmes basés sur les LLM dans les environnements opérationnels (étapes 10 à 11). L'évolution du paysage des techniques d'intégration et des architectures de systèmes composés (étapes 7 à 11) a permis aux modèles autonomes de fonctionner dans des cadres de plus en plus complexes et agentiques. Cela peut inclure, par exemple :

- Des techniques d'augmentation de données telles que la génération augmentée par récupération (RAG), qui permettent aux modèles d'accéder à des sources de données externes ou propriétaires en réponse aux requêtes des utilisateurs ;
- Des flux de travail agents qui orchestrent plusieurs modèles ou outils pour effectuer des tâches en plusieurs étapes avec des degrés d'autonomie variables.

À des étapes avancées de déploiement, des couches supplémentaires de personnalisation peuvent être introduites. Ces derniers reposent fréquemment sur le traitement intensif des données personnelles pour adapter applications, produits ou services aux utilisateurs individuels. Ces mécanismes peuvent inclure des fonctionnalités de décodage de l'intention de l'utilisateur et de personnalisation des prompts conçues pour déduire les préférences, habitudes ou indices contextuels de l'utilisateur afin d'adapter dynamiquement les réponses du système. Bien que ces développements améliorent la fonctionnalité et l'adaptabilité, ils peuvent également élargir la portée, le volume et la sensibilité du traitement des données personnelles, augmentant ainsi les risques liés à la vie privée et à la protection des données.

### 2.1.1 Différence entre le modèle LLM et le système basé sur LLM

Risques modèle vs. système dans l'écosystème LLM

Avec les applications naissantes d'assistance personnelle basées sur des agents (produits dits *d'intelligence personnelle),* les applications populaires basées sur les LLM reposent fortement sur l'utilisation intensive des données personnelles pour créer des services intuitifs et hautement interactifs. Ces fonctionnalités, bien qu'elles soient bénéfiques pour l'ergonomie et une assistance personnalisée améliorée, soulèvent de sérieuses questions sur la protection des données. En particulier, elles brouillent les frontières entre les phases de formation, de personnalisation et de collecte continue de données, augmentant la difficulté de suivre où et comment les données personnelles sont traitées.

Dans ce contexte, il est essentiel de distinguer les risques associés au modèle lui-même de ceux qui émergent au niveau du système plus large dans lequel le modèle est intégré. Cette

distinction n'est pas purement technique ; Il est fondamental pour concevoir des stratégies efficaces, légales et respectueuses du contexte de gestion des risques liés à la vie privée.

Les risques au niveau modèle découlent de la manière dont le LLM est entraîné, affiné en post-entraînement et architecturé, notamment :

- Ingestion de données personnelles lors de la pré-formation, souvent à partir de webscraping à grande échelle, sans transparence ni fondement légal.
- La mémorisation et la régurgitation d'informations sensibles ou identifiables issues des données d'entraînement, ce qui pourrait violer les principes de minimisation des données et de limitation du stockage.
- Hallucinations ou la génération d'informations personnelles plausibles mais fausses, qui peuvent nuire à la réputation ou à la vie privée des personnes concernées.
- Amplification des biais, où les associations statistiques sous-jacentes reproduisent ou renforcent des schémas injustes ou discriminatoires dans la manière dont les données personnelles sont traitées ou référencées.

Les risques au niveau système apparaissent lorsqu'un LLM est intégré à un environnement applicatif plus large, incluant souvent des API, des interfaces, des plug-ins, des fonctions mémoire, des boucles de rétroaction, des systèmes RAG, des flux de travail agents impliquant l'orchestration de LLM, et des services tiers. Ici, les menaces à la vie privée sont liées non seulement à ce que fait le modèle, mais aussi à la manière dont il est utilisé, et par qui. Exemples de risques incluent :

- Profilage utilisateur persistant via des fonctions mémoire comme vu précédemment.
- Manque de transparence sur la manière dont les données utilisateur sont traitées, partagées ou réutilisées, surtout lorsque les LLM fonctionnent dans des systèmes cloud dynamiques.
- La fuite de données intercontextuelle, où les données utilisateur fournies dans un contexte applicatif sont réutilisées dans un autre (par exemple, par un ajustement fin partagé entre produits).
- Contrôles ou mécanismes de consentement inadéquats pour les utilisateurs, en particulier pour les usages secondaires des données, y compris la personnalisation ou les tests A/B.
- Les fonctionnalités mémoire stockent et exploitent l'historique d'interaction des utilisateurs pour améliorer la continuité, personnaliser les réponses, améliorer l'expérience utilisateur final et soutenir une optimisation ultérieure du modèle ou le développement produit.

À chaque étape, la qualité et la disponibilité des données jouent un rôle crucial. Cependant, il est important de distinguer les risques de vie privée associés aux modèles LLM fondateurs de ceux découlant des systèmes basés sur LLM dans des contextes opérationnels.

### 2.1.2. Cinq étapes fondamentales des processus dynamiques de gestion du cycle de vie et des risques des LLM

Écosystème technologique paysager des systèmes basés sur LLM et des données personnelles

Une cartographie générale de l'écosystème technologique des systèmes basés sur LLM, illustrée à la Figure 1, offre un aperçu orienté vers le cycle de vie et clarifie la distinction fondamentale entre les grands modèles de langage fondamentaux et les systèmes plus larges construits avec et autour d'eux. Une telle cartographie démontre que différents acteurs technologiques et économiques peuvent être impliqués à chaque étape, avec des rôles et des responsabilités distincts. Il souligne également que la nature, la portée et la sensibilité des données traitées, y compris les données personnelles, varient considérablement d'une phase à l'autre.

Les principales étapes abordées incluent :

(1) Création de modèles, où les données d'entraînement sont collectées, pré-traitées, et les modèles sont construits en tenant compte de la confidentialité dès la conception.
(2) Adaptation post-entraînement, où les modèles sont instruits, affinés et transformés en outils d'assistance adaptés aux tâches.
(3) Intégration système, dans laquelle les LLM sont intégrés dans des applications ou services (impliquant souvent un ajustement fin de modèles pré-entraînés) avec des garanties appropriées.
(4) Déploiement opérationnel, faisant référence au déploiement en temps réel du système basé sur le LLM avec des contrôles actifs de surveillance et de gouvernance.
(5) L'interaction utilisateur final, couvrant la manière dont les utilisateurs interagissent avec les systèmes basés sur les LLM et comment les flux de travail autonomes ou les fonctions agentiques sont gérés.

Dans l'écosystème des LLM, le paysage du risque est dynamique, dépendant du contexte et multi-niveaux. Pour être efficace, un cadre de gestion des risques de vie privée pour les LLM doit :

1. Traiter à la fois les couches modèle et système : considérer les risques au niveau du modèle statistique *et* de l'application/système dans lequel il opère.
2. Suivez les risques tout au long du cycle de vie : de la pré-formation au déploiement et au-delà, y compris l'apprentissage continu et l'interaction utilisateur.
3. Suivez la variabilité de la réponse par une évaluation continue : comme les LLM sont probabilistes, et non déterministes[1], certaines violations de la confidentialité ne peuvent survenir qu'après le déploiement (par exemple, par régurgitation, injections promptes ou mauvaise utilisation de la sortie).
4. Intégrer à la fois des contrôles techniques et organisationnels : aucune solution technique unique n'est suffisante ; la gouvernance doit combiner les perspectives ingénierie, juridique et UX.

[1] Les systèmes basés sur des LLM diffèrent fondamentalement des logiciels traditionnels en ce qu'ils ne possèdent pas d'ensemble prédéfini de règles produisant un comportement déterministe où une entrée ne correspond qu'à une seule sortie. La probabilité et la prédiction sont au cœur du comportement non déterministe des LLM, où une entrée possède de nombreuses sorties différentes nécessitant un cadre de gouvernance pour intégrer une couche d'évaluation continue.

Une gestion efficace des risques doit s'aligner sur des principes fondamentaux tels que *la légalité*, *l'équité*, *la transparence*, *la limitation des objectifs* et *la responsabilité*. Ces éléments doivent s'appliquer non seulement au modèle lui-même, mais à l'ensemble de l'écosystème dans lequel le modèle est déployé.

## 2.2 Types de risques pour la vie privée Tout au long du cycle de vie de l'IA et dans l'écosystème LLM

### 2.2.1 Contexte Cartographie des risques de confidentialité dans les systèmes basés

Les LLM et les systèmes basés sur les LLM posent des défis importants à chaque étape de leur cycle de vie : des données d'entraînement compromises et des indications adverses lors de l'inférence, à l'opacité systémique dans le déploiement et la surveillance post-déploiement. Les risques liés à la vie privée se manifestent non seulement par la fuite directe ou la mémorisation de données, mais plus largement par des comportements modéliques qui permettent la manipulation d'identité, l'usurpation d'identité, la désinformation ou un comportement prédictif potentiellement manipulateur. Par exemple, le contenu deepfake ou les comptes d'influenceurs utilisant des identités générées par IA se sont multipliés en ligne, renforçant la difficulté de distinguer les personnes réelles des personas synthétiques.

Les risques liés à la vie privée dans les systèmes basés sur des LLM doivent être compris en lien avec les phases spécifiques du modèle et du cycle de vie du système. Chaque phase présente des profils de risque distincts, des expositions système et des défis de gouvernance qui nécessitent une évaluation contextuelle :

1. Phase 1 : Phase de formation impliquant la collecte et le prétraitement des données d'entraînement

   Les risques découlent de l'ingestion incontrôlée de données personnelles dans les ensembles de données d'entraînement. Les corpus accessibles au public peuvent contenir des informations identifiables, telles que des clés API, des e-mails ou des discussions personnelles sensibles. Sans filtrage adéquat, les modèles peuvent mémoriser puis reproduire ce contenu, remis en cause des principes tels que la minimisation des données, la limitation des objectifs et l'équité selon la Convention 108+.

2. Phase 2 : Instruction post-formation, Affinage

   L'ajustement fin introduit des risques, en particulier lorsque des données sensibles ou spécifiques au domaine sont utilisées sans contrôles suffisants. Il peut également amplifier le biais ou déstabiliser le comportement du modèle. De nombreux déployeurs s'appuient sur des modèles affinés proposés comme des services sans transparence totale sur le processus post-formation, compliquant les évaluations de la confidentialité et la responsabilité.

3. Phase 3 : Intégration et déploiement au niveau système Vulnérabilités dans les orchestrations de tâches chez les agents IA et conception d'API

Les risques vont au-delà du modèle pour inclure les API, les middlewares, les architectures de génération augmentée Retrieval (RAG) ainsi que l'orchestration des agents. Des points de terminaison ou des intégrations mal sécurisés peuvent exposer des données privées. Ces questions nécessitent des principes de sécurité dès la conception et renforcent l'importance des audits architecturaux et des mesures de protection en couches.

4. Phase 4 : Déploiement opérationnel/suivi et adaptation post-déploiement
   La collecte continue de données et les mises à jour des modèles manquent souvent de transparence. Les données de retour peuvent être réutilisées de manière à réintroduire des risques pour la vie privée, et les utilisateurs peuvent ignorer comment leurs entrées sont stockées ou analysées. Les articles 10 de la Convention 108+ et l'article 16 du Traité sur l'IA soulignent l'importance d'une surveillance continue et d'évaluations d'impact rigoureuses.

5. Phase 5 : Inférence, interaction utilisateur dans les applications et IA agente
   Le jailbreaking et l'injection rapide sont des menaces majeures à cette étape. Même des modèles bien protégés peuvent être manipulés pour révéler des informations sensibles ou se comporter de manière inappropriée. L'incapacité à prédire ou retracer de manière cohérente les résultats soulève également des questions de transparence et de responsabilité. L'utilisation intensive des données personnelles pour créer des services intuitifs et hautement interactifs soulève les préoccupations de confidentialité déjà évoquées.

Pourtant, au-delà de ces phases structurées, les LLM introduisent également une catégorie plus large de préjudices systémiques et sociaux, qui nécessitent un autre type de surveillance[2].

### 2.2.2 Risques liés à la vie privée dans l'écosystème LLM des données

Alors que la section précédente a cartographié les risques liés à la vie privée à travers les phases du cycle de vie, cette section se concentre spécifiquement sur les catégories de données personnelles traitées dans l'écosystème des LLM. Comprendre quels types de données sont impliqués est essentiel pour évaluer la portée, la sensibilité et l'impact potentiel des risques liés à la vie privée.

L'écosystème de données entourant les systèmes basés sur les LLM est en couches et dynamique. Différentes formes de données personnelles sont traitées à différents stades du développement, de l'intégration et du déploiement du modèle. Ces catégories varient en origine, sensibilité, identification et implications en matière de gouvernance.

Phase 1 : Données de formation et de pré-entraînement
Les ensembles de données d'entraînement peuvent inclure du contenu en ligne accessible au public, des corpus sous licence, des bases de données propriétaires et d'autres sources agrégées. Ces ensembles de données peuvent contenir :

---

[2] Voir par exemple les implications structurelles décrites dans le projet de note d'orientation sur les implications de l'IA générative pour la liberté d'expression par le Comité MSI-AI du Conseil de l'Europe.https ://www.coe.int/en/web/freedom-expression/msi-ai-committee-of-experts-on-the-impacts-of-generative-artificial-intelligence-for-freedom-of-expression# { %22265382451%22 :[1]}

- Informations personnelles identificables, telles que noms, coordonnées ou affiliations professionnelles ;
- Contenu généré par les utilisateurs, y compris les discussions sur les forums, les publications sur les réseaux sociaux et les commentaires ;
- Des catégories spéciales de données, lorsqu'elles ont été extraites ou incluses sans filtrage adéquat ;
- Informations confidentielles intégrées, telles que des clés API ou des communications personnelles.

À ce stade, les données personnelles sont souvent traitées à grande échelle et sans interaction directe avec les personnes concernées. L'ampleur et l'agrégation de telles données augmentent le risque de rétention involontaire, d'inférence d'attributs sensibles et de transparence réduite quant à la provenance et à la base légale.

Phases 1 & 2 : Données de test et d'évaluation en post-entraînement et ajustement fin
Lors des tests, des benchmarks et du red-teaming, des ensembles de données supplémentaires peuvent être introduits. Cela peut inclure des données personnelles sélectionnées, des données synthétiques issues de personnes réelles, des incitations adverses ou l'évaluation de jeux de données conçus pour mesurer la performance. Bien que souvent considérés comme des artefacts techniques, ces ensembles de données peuvent tout de même contenir des informations personnelles identifiables ou déduibles. Leur utilisation soulève des questions concernant le traitement secondaire, les garanties dans les environnements expérimentaux, et la différenciation entre données anonymisées et données pseudonymisées.

Phases 3 & 4 : Données opérationnelles et interactionnelles
Une fois déployés, les systèmes basés sur des LLM traitent les données générées par l'interaction utilisateur. Celles-ci peuvent inclure :

- Requêtes et prompts des utilisateurs ;
- Documents téléchargés ou entrées contextuelles ;
- Contenu récupéré dans des architectures telles que les systèmes augmentés par récupération ;
- Journaux d'utilisation, télémétrie et métadonnées comportementales.

Les données opérationnelles sont souvent très contextuelles et peuvent révéler des préférences, des intentions, des habitudes ou des circonstances personnelles sensibles. Contrairement aux données d'entraînement, ces entrées sont directement liées à des utilisateurs identifiables et peuvent être traitées en continu ou conservées pour l'optimisation du système.

Phase 5 : Surveillance, retour d'information et réentraînement des données
Les mécanismes d'amélioration post-déploiement reposent fréquemment sur des données de rétroaction, des journaux de performance et des enregistrements d'interaction sélectionnés. Au fil du temps, ces boucles de rétroaction peuvent générer de nouveaux ensembles de données dérivés du comportement des utilisateurs. Un tel traitement itératif des données peut brouiller les frontières entre les objectifs initiaux et l'optimisation

ultérieure, soulevant des préoccupations liées à la limitation des objectifs, aux périodes de conservation et à l'allocation des rôles entre les acteurs de l'écosystème.

Phase 5 : Archivage et Données de suppression
À des étapes ultérieures, des décisions doivent être prises concernant la conservation, l'anonymisation ou la suppression des données d'entraînement, des journaux opérationnels et des ensembles de données dérivés. Dans les écosystèmes multi-acteurs, déterminer la responsabilité de l'effacement ou de la rectification peut être complexe, en particulier lorsque des données personnelles sont intégrées dans des paramètres de modèles, des journaux ou des infrastructures distribuées.

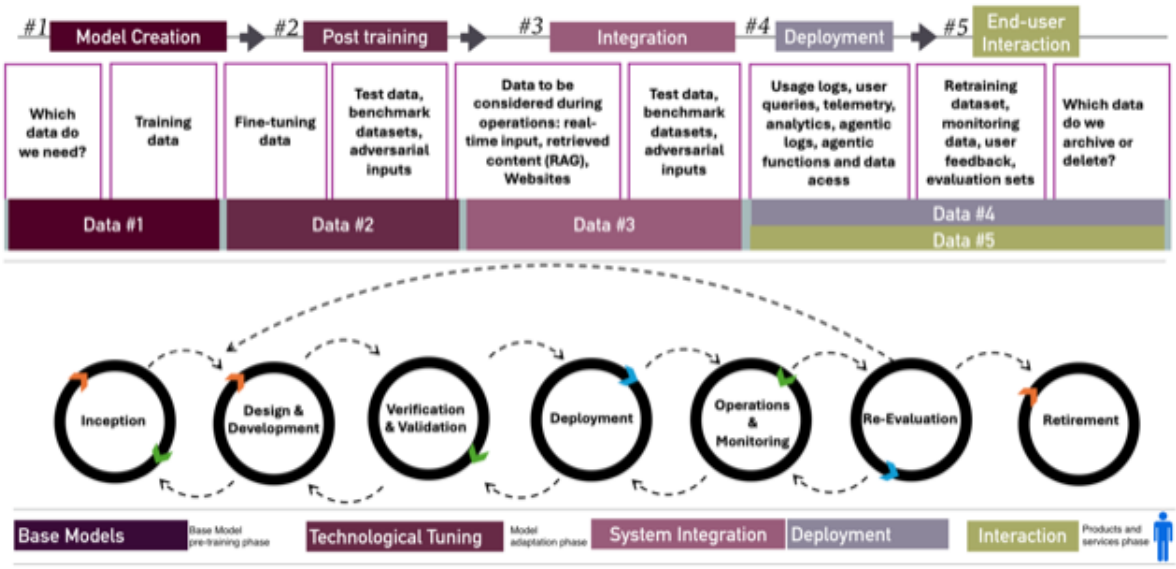


*Figure 2 : Illustre la répartition des risques liés à la vie privée à différents stades du cycle de vie des systèmes basés sur des LLM, en mettant l'accent sur l'accès et les flux aux données. (Modifié à partir du rapport de l'EDPB sur les risques et atténuations de la vie privée dans les LLM. Les phases du cycle de vie de l'ISO/IEC 22989 sont également illustrées sur l'image.*

## 2.3 Risques émergents pour la vie privée

À mesure que les LLM sont de plus en plus intégrés dans les infrastructures publiques et privées, l'absence de garanties adaptées en matière de confidentialité est devenue une préoccupation croissante. Dans le cadre de la Convention 108+, ces systèmes introduisent à la fois des risques nouveaux et systémiques, allant de l'exposition de données personnelles à l'érosion de la vie privée par l'inférence opaque, les prédictions et le profilage.

Les pratiques actuelles ne sont pas bien équipées pour faire face à ces risques. Les évaluations de risques traditionnelles ne parviennent souvent pas à saisir toute l'étendue et l'imprévisibilité des modèles génératifs comme les LLM et les systèmes composites et architectures agentiques sur lesquels ils s'appuient. De nombreuses organisations n'ont pas une vision claire de l'origine de leurs données de formation ni de la manière dont les usages en aval pourraient impacter les droits individuels. En même temps, les LLM brouillent la frontière entre informations synthétiques et informations personnelles. Leurs résultats peuvent inclure des données mémorisées ou du contenu sensible généré en réponse à des consignes ingénieusement conçues, ce qui rend difficile la détection des problèmes avant le déploiement.

Ces risques vont au-delà de la protection des données. Les LLM influencent également notre compréhension de l'autonomie et de l'identité personnelles. À mesure que le contenu généré par machine devient plus difficile à distinguer de la communication humaine, les gens peuvent avoir du mal à prouver ce qu'ils ont dit ou non. Cette érosion de l'auteur et de l'authenticité soulève des préoccupations plus profondes ; Pas seulement de la vie privée, mais aussi de la dignité, de la confiance et de la tension psychologique d'interagir avec des systèmes capables d'imiter les individus de façon si convaincante. Ces préjudices remettent en cause des aspects

fondamentaux de l'identité individuelle et de l'autonomie, affectant non seulement les garanties de la Convention 108+, mais aussi les protections consacrées à l'article 8 de la Convention européenne des droits de l'homme (CEDH), qui garantit le droit à l'identité, à la réputation et à la vie privée.