AI Act GPAI 2026 : Obligations des Modèles d'IA à Usage Général

Guide AI Act 2026 pour modèles GPAI — obligations transparence, évaluation risques systémiques, droits d'auteur et sanctions pour fournisseurs IA en France.

L'entrée en vigueur progressive du règlement européen sur l'intelligence artificielle (AI Act) marque un tournant historique pour les fournisseurs de modèles d'IA à usage général — les **GPAI (General Purpose AI)**. À compter d'août 2025 pour les premières obligations et jusqu'en 2026 pour le régime complet, les acteurs qui développent ou distribuent ces modèles massifs font face à un cadre réglementaire inédit, à la fois ambitieux et techniquement exigeant. En France, la CNIL, en lien étroit avec l'AI Office européen nouvellement créé, est chargée de la supervision nationale. Les entreprises qui ignorent ces obligations s'exposent à des amendes pouvant atteindre 3 % du chiffre d'affaires mondial ou 15 millions d'euros. Ce guide technique décrypte l'ensemble des obligations GPAI applicables en 2026 : définition légale, seuil systémique, documentation technique, résumé droits d'auteur, signalement d'incidents, Code de pratique, articulation avec ISO 42001 et calendrier d'application. Que vous soyez fournisseur de LLM, intégrateur d'API ou opérateur d'un système IA basé sur un modèle fondateur, ce référentiel vous donnera les clés pour construire une conformité solide et durable face aux exigences de l'AI Office et des autorités nationales de contrôle.

À RETENIR

À retenir — GPAI & AI Act 2026

• Seuil systémique : 10^{25} FLOPs d'entraînement (présomption réfragable)

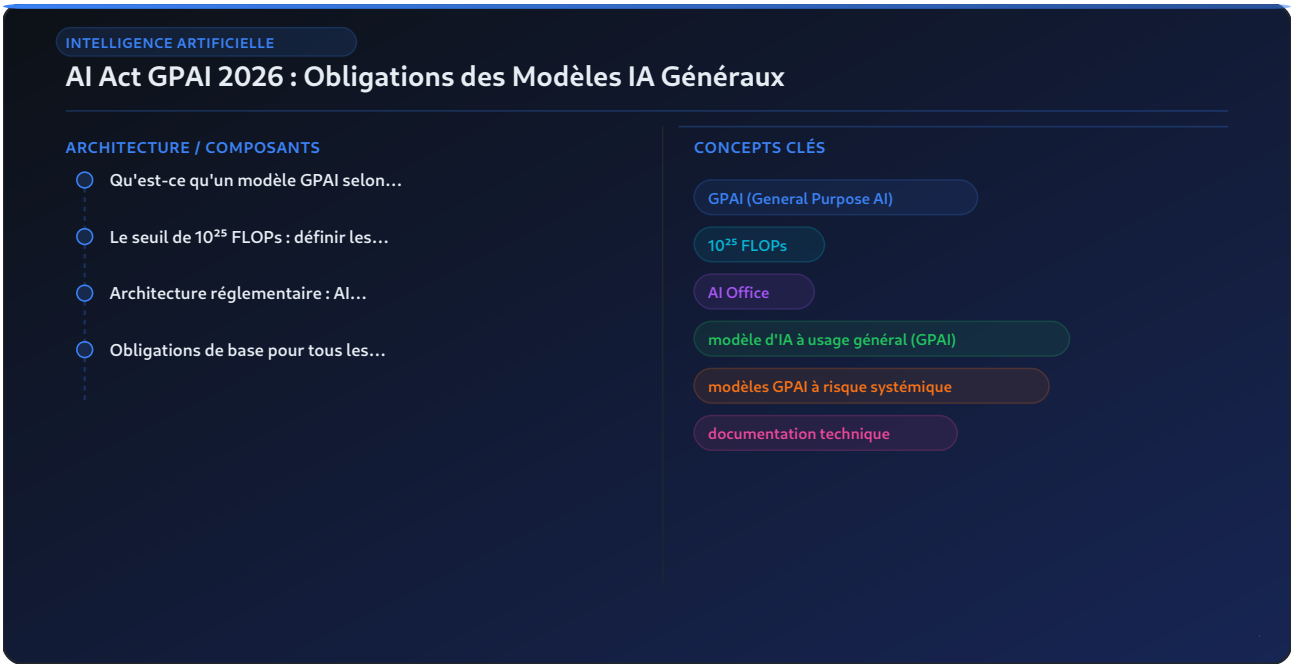
• Obligations de base GPAI : documentation technique, politique droits d'auteur, résumé données

• Obligations systémiques supplémentaires : évaluation adversariale, signalement incidents graves

• L'**AI Office** (Commission européenne) est l'autorité de supervision principale

• Sanctions : jusqu'à 3 % CA mondial ou 15 M€ (violations GPAI)

• Code de pratique GPAI : première version finale publiée mi-2025, adhésion recommandée



Qu'est-ce qu'un modèle GPAI selon l'AI Act ?

L'AI Act définit le **modèle d'IA à usage général (GPAI)** comme un modèle entraîné sur une grande quantité de données, capable d'accomplir une large gamme de tâches distinctes et déployable dans de multiples systèmes ou applications. Cette définition volontairement large englobe les grands modèles de langage (LLM), les modèles multimodaux (texte+image+audio), les modèles de génération d'image et les modèles de code. L'article 3(63) du règlement précise que le caractère "usage général" s'apprécie par la polyvalence intrinsèque du modèle, indépendamment de l'usage effectif fait par le déployeur. Un modèle fine-tuné reste GPAI si le

modèle de base l'est. La distinction avec les systèmes IA "à usage spécifique" est fondamentale : un modèle de détection de fraude entraîné uniquement sur des données bancaires n'est pas GPAI. En revanche, GPT-4, Claude 3, Gemini Ultra ou Mistral Large entrent pleinement dans cette catégorie. Pour les fournisseurs en aval (intégrateurs API), le règlement prévoit un régime de responsabilité partagée avec le fournisseur du modèle de base.

Le seuil de 10^{25} FLOPs : définir les modèles systémiques

L'AI Act introduit une sous-catégorie critique : les **modèles GPAI à risque systémique**. Le critère quantitatif principal repose sur la puissance de calcul utilisée pour l'entraînement, exprimée en *FLOPs (Floating Point Operations)*. Le seuil fixé par le règlement est de **10^{25} FLOPs** cumulés sur l'entraînement. En 2026, ce seuil correspond approximativement aux modèles de la taille de GPT-4 ou Gemini Ultra. Ce seuil n'est pas figé : la Commission européenne peut l'adapter par acte délégué à mesure que la puissance de calcul accessible augmente. La présomption est *réfragable* : un fournisseur peut démontrer qu'un modèle dépassant le seuil ne présente pas de risque systémique en raison de ses caractéristiques particulières (architecture, domaine de déploiement, mesures de sécurité intégrées). À l'inverse, l'AI Office peut qualifier de systémique un modèle sous le seuil si des "capabilities" particulières sont identifiées — notamment la capacité à mener des cyberattaques sophistiquées, à générer des contenus CSAM, ou à induire des risques pour les infrastructures critiques.

⚡ **Retour terrain**

J'ai audité le pipeline de génération de contenu d'un pure-player e-commerce qui produisait 500 fiches produit/jour avec GPT-4. Le taux de doublons sémantiques (deux produits différents avec des descriptions quasi-identiques) atteignait 12 %. L'ajout d'une

étape de vérification par similarité cosinus avant publication a ramené ce taux à 0,8 % — et le SEO du site a progressé de façon mesurable le mois suivant.

Architecture réglementaire : AI Office, CNIL et autorités nationales

La gouvernance de l'AI Act repose sur une architecture à deux niveaux. Au niveau européen, l'**AI Office** — rattaché à la Commission européenne et opérationnel depuis mai 2024 — détient la compétence exclusive de supervision des modèles GPAI, y compris les modèles systémiques. Il publie les lignes directrices, évalue la conformité des Codes de pratique, et peut mener des enquêtes. Au niveau national, en France, la **CNIL** joue un rôle de coordination avec l'AI Office et de surveillance des systèmes IA déployés sur le territoire. La CNIL dispose également d'attributions propres sur les questions de protection des données dans les systèmes IA (articulation RGPD-AI Act). D'autres autorités sectorielles interviennent selon le domaine : l'AMF pour la finance, l'ANSM pour la santé, l'ARCEP pour les télécoms. Cette fragmentation crée un risque de "forum shopping" réglementaire, mais l'AI Office garde le dernier mot sur les questions purement GPAI.

Obligations de base pour tous les fournisseurs GPAI

L'article 53 de l'AI Act liste les obligations s'appliquant à *tous* les fournisseurs de modèles GPAI, quel que soit le seuil de FLOPs. Ces obligations de base incluent : la rédaction et le maintien d'une **documentation technique** couvrant l'architecture du modèle, les données d'entraînement, les capacités et limitations, et les performances sur les benchmarks standards ; l'établissement d'une **politique de conformité au droit d'auteur** documentant les mesures prises pour respecter les droits des titulaires lors de la constitution des données d'entraînement (conformément à la directive 2019/790) ; la publication d'un **résumé des données d'entraînement** suffisamment détaillé pour permettre à des tiers de comprendre les sources utilisées ; et la mise à disposition

d'informations techniques aux opérateurs en aval pour leur permettre de respecter leurs propres obligations. Ces obligations sont applicables depuis août 2025 pour les nouveaux modèles, avec une période de transition pour les modèles déjà disponibles avant cette date.

Documentation technique GPAI : contenu et format attendus

La documentation technique requise par l'AI Act n'est pas un simple datasheet marketing : elle doit être suffisamment précise pour permettre à l'AI Office d'évaluer la conformité du modèle. L'annexe XI du règlement détaille le contenu minimal. En pratique, un dossier de documentation GPAI complet doit inclure : la description de l'architecture (nombre de paramètres, type de modèle, modalités) ; les sources et volumétrie des données d'entraînement (y compris les jeux de données tiers et leur licence) ; la méthodologie de filtrage et de nettoyage des données ; les performances mesurées sur des benchmarks standards (MMLU, HumanEval, TruthfulQA…) ; les capacités identifiées et les "misuse risks" documentés ; les mesures techniques de sécurité intégrées (alignement, RLHF, constitutional AI…) ; et les restrictions d'utilisation (use-case policies). Cette documentation doit être mise à jour à chaque version significative du modèle et conservée pendant 10 ans après la fin de la mise sur le marché.

Droits d'auteur et résumé des données d'entraînement

L'une des obligations les plus sensibles pour les fournisseurs GPAI concerne la **transparence sur les données d'entraînement** au regard du droit d'auteur. L'article 53(1)(d) exige que les fournisseurs publient un résumé "suffisamment détaillé" des données utilisées, incluant les sources web crawlées, les bases de données licenciées, les contenus générés par l'humain et les données synthétiques. En parallèle, la politique de conformité droit d'auteur doit documenter les mécanismes mis en place pour respecter les opt-outs prévus par la directive copyright (article 4) — notamment via le protocole **TDM Reservation (robots.txt + meta tags)**. Un retour terrain : lors d'un audit GPAI conduit en France début 2026, nous avons constaté que la majorité des

fournisseurs européens de taille intermédiaire ne disposaient pas encore d'un inventaire précis de leurs sources de crawl, rendant impossible la rédaction d'un résumé conforme. La constitution de cet inventaire rétroactif est la tâche la plus chronophage de la mise en conformité GPAI.

Obligations supplémentaires pour les modèles GPAI systémiques

Les modèles franchissant le seuil des 10^{25} FLOPs (ou qualifiés systémiques par l'AI Office) sont soumis à des obligations additionnelles listées à l'article 55. Premièrement, l'**évaluation adversariale des modèles (red-teaming)** : le fournisseur doit réaliser des tests de robustesse, de sécurité et d'alignement, potentiellement avec des tiers indépendants accrédités. Deuxièmement, le **signalement des incidents graves** à l'AI Office dans un délai maximal de 15 jours ouvrables après connaissance — un incident grave étant défini comme un risque grave pour la santé, la sécurité ou les droits fondamentaux. Troisièmement, la mise en place de **mesures de cybersécurité adéquates** pour protéger le modèle lui-même (model stealing, extraction d'informations sensibles des données d'entraînement via prompt injection…). Quatrièmement, le rapport annuel de consommation énergétique à l'AI Office. Ces obligations systémiques s'appliquent depuis août 2025.

Le Code de pratique GPAI : adhésion et contenu

L'AI Act prévoit la création d'un **Code de pratique GPAI** co-élaboré par les fournisseurs, la société civile et des experts, sous la coordination de l'AI Office. Ce Code constitue le principal moyen de démontrer la conformité aux obligations GPAI : un fournisseur qui respecte le Code bénéficie d'une **présomption de conformité**. La première version finale du Code a été publiée mi-2025 après plusieurs phases de consultation. Il couvre notamment les standards de documentation technique, les méthodologies d'évaluation adversariale, les protocoles de signalement d'incidents et les bonnes pratiques pour les données d'entraînement. Son adhésion est formellement volontaire, mais en pratique quasi-indispensable pour les acteurs

commercialisant leurs modèles en Europe. Les fournisseurs n'adhérant pas au Code doivent démontrer par d'autres moyens leur conformité — une charge de preuve considérablement plus lourde face à l'AI Office.

Évaluation des risques systémiques : méthodologie pratique

L'évaluation des risques systémiques pour les modèles GPAI constitue une nouveauté absolue par rapport aux frameworks existants. Elle combine des approches quantitatives (benchmarks de capabilities, évaluations WMDP pour les risques CBRN) et qualitatives (scénarios d'usage abusif, analyse d'impacts en cascade). Les **domaines de risque prioritaires** identifiés par l'AI Office incluent : les capacités de cyberattaque offensive (génération d'exploits, aide à la conception de malwares) ; les risques CBRN (Chemical, Biological, Radiological, Nuclear) liés à l'aide à la synthèse de substances dangereuses ; les risques pour les processus démocratiques (génération massive de désinformation) ; et les risques pour les infrastructures critiques. La méthodologie recommandée suit une approche en quatre étapes : identification des capabilities dangereuses, évaluation de l'accessibilité de ces capabilities, estimation de l'impact potentiel, et proposition de mesures d'atténuation. Des organismes comme Apollo Research ou UK AI Safety Institute ont développé des protocoles d'évaluation de référence.

Articulation AI Act et ISO/IEC 42001

La norme **ISO/IEC 42001:2023** — premier système de management de l'IA certifiable — offre un cadre complémentaire à l'AI Act pour les fournisseurs GPAI. Bien que la certification ISO 42001 ne constitue pas automatiquement une présomption de conformité à l'AI Act, les deux référentiels sont largement alignés sur plusieurs points clés. La gestion des risques IA (clause 6.1.2 de l'ISO 42001) couvre les mêmes domaines que l'évaluation des risques systémiques AI Act. Le système de documentation requis (clause 7.5) répond en grande partie aux exigences de l'annexe XI. La gestion des fournisseurs (clause 8.4) s'applique aux relations avec les

fournisseurs de données d'entraînement et les intégrateurs. En pratique, les entreprises qui ont initié une démarche ISO 42001 disposent d'une avance significative pour la conformité AI Act. L'écart principal concerne les obligations de signalement spécifiques aux modèles systémiques, qui n'ont pas d'équivalent direct dans la norme ISO.

Obligations en cascade : fournisseurs et opérateurs en aval

L'AI Act crée un système de **responsabilité en cascade** particulièrement important pour les modèles GPAI distribués via API. Le fournisseur du modèle de base a l'obligation de transmettre aux opérateurs en aval toutes les informations nécessaires pour que ces derniers puissent respecter leurs propres obligations. En pratique, cela se traduit par la mise à disposition d'une **model card** détaillée, d'une documentation des limitations et risques identifiés, et d'une politique d'utilisation acceptable (AUP). Les opérateurs qui intègrent un modèle GPAI dans un système IA à haut risque (catégorie 8 de l'AI Act) restent responsables de la conformité du système final, mais peuvent s'appuyer sur la documentation du fournisseur. La chaîne de responsabilité doit être formalisée contractuellement — les clauses "IA" dans les contrats cloud et API deviennent un enjeu juridique majeur en 2026.

Signalement des incidents graves : procédure et délais

Le régime de **signalement d'incidents graves** prévu pour les modèles GPAI systémiques s'inspire du RGPD (notification de violations de données) tout en présentant des spécificités importantes. Un incident grave au sens de l'AI Act inclut tout événement ayant causé ou susceptible de causer un préjudice grave à des personnes ou à des groupes de personnes, ou ayant compromis la sécurité ou l'intégrité du modèle de manière significative. Le délai de notification à l'AI Office est de **15 jours ouvrables** après la découverte de l'incident. La notification initiale peut être partielle si l'investigation n'est pas terminée, mais doit inclure a minima : la nature de l'incident, son impact estimé, les mesures d'atténuation immédiate prises, et le plan

d'investigation. Un rapport final complet est attendu dans les 30 jours supplémentaires. Les fournisseurs doivent maintenir un **registre des incidents** incluant également les incidents non-graves, disponible à l'inspection de l'AI Office sur demande.

Cybersécurité des modèles GPAI : vecteurs d'attaque spécifiques

Les modèles GPAI présentent une surface d'attaque spécifique que les mesures de cybersécurité traditionnelles ne couvrent pas. L'AI Act exige des fournisseurs systémiques qu'ils mettent en place des mesures adéquates contre les **vecteurs d'attaque propres aux LLM**. Le premier vecteur est l'**extraction d'informations d'entraînement** : des techniques de membership inference ou de model inversion permettent potentiellement de reconstruire des données sensibles du corpus d'entraînement. Le second est le **model stealing** : via un grand nombre de requêtes API, un attaquant peut distiller un modèle approximatif du modèle commercial. Le troisième, et le plus immédiat, est l'**injection de prompt adversariale** : des instructions malveillantes intégrées dans le contexte peuvent forcer le modèle à outrepasser ses garde-fous. Les mesures de mitigation incluent : rate limiting côté API, détection d'anomalies dans les patterns de requête, techniques de differential privacy lors de l'entraînement, et red-teaming régulier des mécanismes de sécurité.

Sanctions AI Act pour les violations GPAI

Le régime de sanctions de l'AI Act distingue plusieurs niveaux selon la gravité de la violation. Pour les manquements aux obligations GPAI (documentation, politique droits d'auteur, résumé données, signalement incidents), l'amende maximale est de **3 % du chiffre d'affaires mondial annuel** de l'entreprise ou **15 millions d'euros**, le montant le plus élevé s'appliquant. Pour les violations des obligations des modèles systémiques (évaluation adversariale, mesures de cybersécurité), le même plafond s'applique. Les sanctions les plus lourdes — jusqu'à **7 % du CA mondial ou 35 M€** — concernent les violations des interdictions absolues (systèmes IA

interdits) qui ne s'appliquent pas directement aux modèles GPAI en tant que tels. La fourniture d'informations incorrectes à l'AI Office ou aux autorités nationales est sanctionnée à hauteur de **1 % du CA mondial ou 7,5 M€**. Pour les PME et start-ups, des plafonds réduits s'appliquent, calculés sur la base des revenus annuels de l'entreprise concernée.

Timeline d'application : ce qui est applicable en 2026

L'AI Act suit un calendrier d'application progressif depuis son entrée en vigueur en août 2024. En **février 2025**, les interdictions absolues (article 5) sont devenues applicables — manipulation subliminale, notation sociale généralisée, police prédictive basée sur profilage pur. En **août 2025**, les obligations GPAI de base (article 53) et systémiques (article 55) sont entrées en application, ainsi que les règles de gouvernance (chapitre VI). En **août 2026**, l'ensemble du règlement sera pleinement applicable, incluant les exigences pour les systèmes IA à haut risque (chapitres III et IV) et les règles relatives à certains systèmes IA à usage général intégrés dans des produits couverts par d'autres législations européennes. Les fournisseurs GPAI qui ont mis leurs modèles sur le marché *avant* août 2025 bénéficient d'un régime transitoire jusqu'en août 2027 pour les obligations de documentation, mais les obligations de signalement d'incidents s'appliquent dès août 2025.

Mise en conformité pratique : roadmap pour les fournisseurs GPAI

Une roadmap de conformité GPAI réaliste pour 2026 doit être structurée en quatre grandes phases. La **phase 1 (inventaire et gap analysis)** consiste à cartographier tous les modèles GPAI de l'entreprise, identifier ceux dépassant le seuil systémique, et évaluer les écarts par rapport aux obligations légales — notamment la documentation technique et le résumé des données d'entraînement. La **phase 2 (documentation)** couvre la rédaction de la documentation technique (annexe XI), la politique droits d'auteur, et le résumé des données, ainsi que la mise à jour des model cards et des contrats avec les opérateurs en aval. La **phase 3 (évaluation adversariale)**

implique la mise en place ou le renforcement des processus de red-teaming, l'exécution des évaluations de capabilities dangereuses, et la documentation des résultats pour l'AI Office. La **phase 4 (organisation interne)** comprend la désignation d'un responsable conformité AI Act, la création du registre des incidents, la mise en place des procédures de signalement, et l'adhésion formelle au Code de pratique GPAI.

Relation avec le RGPD : les zones de friction

L'articulation entre l'AI Act et le **RGPD** crée plusieurs zones de friction pratiques pour les fournisseurs GPAI. La première concerne les **données personnelles dans le corpus d'entraînement** : le résumé des données requis par l'AI Act peut révéler l'utilisation de données personnelles, ce qui déclenche des obligations RGPD (base légale, droits des personnes concernées). La CNIL a publié des recommandations sur ce point, soulignant la nécessité d'une base légale solide (intérêt légitime ou mission d'intérêt public) pour les usages légitimes de crawl de données. La deuxième friction concerne les **droits des individus** : une personne dont les données figurent dans le corpus d'entraînement peut exercer son droit d'effacement RGPD, mais les modèles de machine learning ne permettent pas l'effacement granulaire sans réentraînement complet — une contrainte technique que ni le RGPD ni l'AI Act ne résolvent pleinement. La troisième friction est l'**évaluation d'impact (DPIA vs. évaluation risques AI Act)** : les deux exercices peuvent être fusionnés en pratique mais obéissent à des logiques différentes.

Cas pratique : conformité d'un LLM B2B en France

Considérons le cas d'une entreprise française développant un LLM spécialisé en droit et fiscalité, commercialisé via API auprès de cabinets d'avocats et d'experts-comptables. Ce modèle, entraîné sur des corpus juridiques sous licence et des textes officiels (Légifrance, BOFIP), compte 70 milliards de paramètres et a nécessité environ 10^{23} FLOPs d'entraînement — sous le seuil systémique. Il est néanmoins soumis aux obligations GPAI de base. La **documentation**

technique devra décrire l'architecture, les sources licenciées utilisées, les benchmarks juridiques (LegalBench, BarExam) et les limitations identifiées (hallucinations jurisprudentielles). La **politique droits d'auteur** devra documenter les licences obtenues auprès des éditeurs juridiques. Le **résumé des données** pourra mentionner les bases légales publiques et les accords avec les éditeurs. Vis-à-vis des cabinets clients (opérateurs), l'entreprise devra fournir model card, AUP et documentation des limitations — notamment l'absence de capacité de conseil personnalisé et la nécessité d'une supervision humaine.

Open source et GPAI : un régime allégé

L'AI Act prévoit un régime spécifique pour les **modèles GPAI open source** — une concession majeure obtenue lors des négociations finales pour protéger l'écosystème open source européen. Les fournisseurs de modèles GPAI dont les poids sont mis à disposition publiquement sous licence libre (MIT, Apache 2.0, Llama community license…) sont exemptés des obligations de documentation technique et de résumé des données vis-à-vis du public général. Ils restent toutefois soumis à la politique de conformité droits d'auteur. L'exemption **ne s'applique pas** aux modèles open source systémiques (dépassant 10^{25} FLOPs) : Mistral Large, Llama 3 70B et leurs successeurs demeurent soumis aux obligations systémiques si le seuil est franchi. Cette asymétrie a fait l'objet de vives critiques de la part des partisans du logiciel libre, qui arguent qu'elle crée une discrimination injustifiée entre modèles open source légers et lourds.

Rôle des évaluateurs tiers et organismes notifiés

Pour les modèles GPAI systémiques, l'AI Act permet à l'AI Office de mandater des **évaluations indépendantes** par des tiers qualifiés — une nouveauté par rapport au régime d'auto-déclaration habituel. Ces évaluateurs tiers, dont la liste sera établie par la Commission, pourront accéder aux modèles, aux données d'entraînement et à la documentation interne pour conduire leurs évaluations. En parallèle, les **organismes notifiés** (Notified Bodies) jouent un rôle pour les

systèmes IA à haut risque, mais pas directement pour les modèles GPAI en tant que tels. L'émergence d'un marché d'évaluation de modèles GPAI est une tendance forte de 2025-2026, avec des acteurs spécialisés comme Apollo Research (UK), AISIs (US/UK) ou le futur Centre d'Évaluation de l'AI Office européen. Pour les fournisseurs, préparer l'accès externe à leurs modèles — dans un cadre contractuel protégeant la confidentialité industrielle — est une anticipation nécessaire.

Impact sur les start-ups françaises et l'écosystème IA

La France dispose d'un écosystème IA dynamique — Mistral AI, Poolside, Nabla, Dust, Alan — qui sera directement impacté par les obligations GPAI. La charge de conformité est proportionnellement plus lourde pour les start-ups que pour les grands groupes, qui disposent d'équipes juridiques dédiées. L'opinion tranchée que nous assumons : l'AI Act GPAI est globalement bien calibré pour les risques systémiques, mais **sous-estime la charge administrative pour les acteurs européens de taille intermédiaire**, créant un avantage concurrentiel indirect pour les hyperscalers américains et les acteurs asiatiques qui peuvent amortir les coûts de conformité sur des revenus massifs. Le dispositif de soutien prévu (sandboxes réglementaires, accompagnement PME par les autorités nationales) est insuffisant face à l'ampleur des obligations documentaires. Une rationalisation via des modèles-types de documentation technique serait bienvenue et constitue une demande forte de l'écosystème français à l'AI Office.

Surveillance continue et mises à jour du modèle

La conformité AI Act n'est pas un exercice ponctuel mais un processus continu. Chaque **mise à jour significative** d'un modèle GPAI déclenche une obligation de mise à jour de la documentation technique et potentiellement une nouvelle évaluation des risques. L'AI Act ne définit pas précisément ce qu'est une "mise à jour significative" pour les modèles GPAI, laissant

une marge d'interprétation. Le Code de pratique GPAI comble partiellement ce vide en proposant des critères pratiques : changement de plus de X% des paramètres, extension significative des capabilities, modification des données d'entraînement. Les fournisseurs doivent également maintenir une **veille réglementaire** active, l'AI Office ayant le pouvoir de publier des lignes directrices contraignantes et des actes délégués modifiant certains aspects du règlement (notamment le seuil systémique et les listes d'usages interdits).

Ressources et liens utiles pour la conformité GPAI

L'écosystème de ressources pour la conformité GPAI s'est considérablement structuré en 2025-2026. L'**AI Office** publie régulièrement des guidelines, des FAQ et les versions du Code de pratique sur son site officiel. La **CNIL** a mis en ligne un guide pratique AI Act pour les acteurs français, couvrant l'articulation RGPD-AI Act. Le **MEDEF** et **France Digitale** ont constitué des groupes de travail produisant des guides sectoriels. Des cabinets spécialisés (Allen & Overy, Bird & Bird, Gide) ont développé des checklist de conformité GPAI. Côté technique, les dépôts GitHub de ML Commons (MLPerf), NIST (AI RMF) et MITRE (ATLAS) fournissent des ressources pour la documentation des capabilities et l'évaluation des risques. Pour les audits techniques, des outils open source comme **LM-Evaluation-Harness** permettent de reproduire les benchmarks standards requis par la documentation technique.

Perspectives : vers un AI Act 2.0 dès 2027 ?

L'AI Act prévoit une clause de révision en 2027, permettant à la Commission d'adapter le règlement à l'évolution rapide de la technologie. Plusieurs aspects sont déjà identifiés comme nécessitant une révision. Le **seuil de 10^{25} FLOPs** risque d'être rapidement obsolète si l'efficacité computationnelle continue de progresser à son rythme actuel — des modèles très puissants pourront bientôt être entraînés avec moins de FLOPs. La **définition des risques systémiques** devra être affinée en fonction des retours d'expérience des premières évaluations adversariales.

Le **régime de responsabilité en cascade** pour les systèmes IA construits sur des GPAI devra être clarifié, notamment pour les systèmes agentiques multi-modèles. La perspective d'un AI Act 2.0 ne doit pas conduire à une attentisme : les obligations actuelles sont applicables et les autorités européennes ont clairement signalé leur intention de commencer les contrôles dès 2026.

FAQ — AI Act GPAI 2026

Mon entreprise utilise l'API de GPT-4 pour un chatbot interne — est-elle soumise aux obligations GPAI ?

Non, directement. Si vous utilisez l'API d'OpenAI (ou d'un autre fournisseur GPAI), c'est OpenAI qui est soumis aux obligations de fournisseur GPAI. Votre entreprise est un *opérateur* au sens de l'AI Act. Vous avez des obligations d'opérateur — notamment la transparence vis-à-vis des utilisateurs si le système prend des décisions qui les affectent significativement — mais pas les obligations de documentation technique ni de résumé des données propres aux fournisseurs GPAI.

Un modèle fine-tuné sur GPT-4 est-il considéré comme GPAI ?

Oui, si le fine-tuning maintient les capacités générales du modèle de base. L'AI Act considère que le fournisseur du modèle fine-tuné est un nouveau fournisseur GPAI et assume les obligations correspondantes. Il peut toutefois s'appuyer sur la documentation du modèle de base (fournie par OpenAI) pour constituer sa propre documentation, en y ajoutant les spécificités liées au fine-tuning (données utilisées, capabilities modifiées).

Quel est le délai pour se mettre en conformité si mon modèle a été mis sur le marché avant août 2025 ?

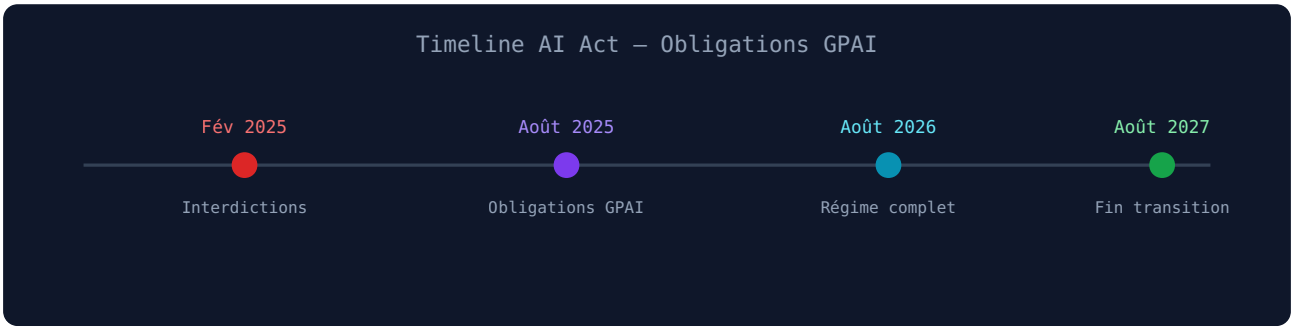
Pour les modèles GPAI mis sur le marché avant août 2025, une période de transition s'applique jusqu'en août 2027 pour les obligations de documentation technique (article 53). Toutefois, les

obligations de signalement d'incidents graves pour les modèles systémiques sont applicables depuis août 2025, sans période de transition. Il est donc impératif de mettre en place les procédures de détection et de signalement d'incidents dès maintenant, même si la documentation complète n'est pas encore finalisée.

Comment prouver la conformité à l'AI Act sans certification officielle pour les GPAI ?

L'adhésion au Code de pratique GPAI crée une présomption de conformité. À défaut, le fournisseur doit documenter son approche alternative et la mettre à disposition de l'AI Office sur demande. Des audits techniques par des tiers qualifiés peuvent renforcer la crédibilité de la démarche. La tenue d'un registre détaillé de toutes les actions de conformité — avec dates, responsables et preuves — est indispensable pour pouvoir démontrer la bonne foi en cas d'enquête.

Obligation	GPAI Standard	GPAI Systémique	Applicable depuis
Documentation technique (Annexe XI)	√ Oui	√ Oui (étendue)	Août 2025
Politique conformité droits d'auteur	√ Oui	√ Oui	Août 2025
Résumé données d'entraînement	√ Oui	√ Oui	Août 2025
Évaluation adversariale (red-teaming)	✗ Non	√ Oui	Août 2025
Signalement incidents graves (15j)	✗ Non	√ Oui	Août 2025
Rapport consommation énergétique	✗ Non	√ Oui	Août 2025
Mesures cybersécurité modèle	✗ Non	√ Oui	Août 2025
Enregistrement auprès de l'AI Office	√ Oui	√ Oui	Août 2026



Red-teaming des modèles GPAI : protocole et bonnes pratiques

Le **red-teaming** des modèles GPAI, rendu obligatoire pour les modèles systémiques, est une discipline qui dépasse largement le test de pénétration classique. Il s'agit d'une évaluation adversariale systématique visant à identifier les capabilities dangereuses du modèle avant son déploiement public. Un protocole de red-teaming GPAI efficace se décompose en plusieurs phases distinctes. La **phase de cartographie des capabilities** identifie toutes les fonctionnalités avancées du modèle via des prompts exploratoires — rédaction d'exploits, synthèse de substances dangereuses, génération de désinformation ciblée. La **phase d'évaluation des barrières** teste la robustesse des mécanismes de sécurité intégrés face aux techniques de jailbreak connues : prompt injection, many-shot jailbreaking, role-playing adversarial, encodage alternatif (base64, langue étrangère). La **phase de test de persistance** évalue si les comportements dangereux peuvent être induits via des conversations longues où les garde-fous s'affaiblissent progressivement. Pour les modèles systémiques, l'AI Office recommande d'inclure des red-teamers externes issus de domaines techniques spécifiques (cybersécurité, chimie, biologie) pour tester les capabilities dans leurs domaines d'expertise. Les résultats — y compris les vulnerabilités non corrigées avec justification — doivent figurer dans la documentation technique. Cette transparence partielle avec l'AI Office vise à permettre la détection de patterns communs entre modèles, contribuant à la sécurité collective de l'écosystème GPAI européen. En France, des équipes spécialisées dans l'évaluation adversariale de LLM commencent à émerger, souvent issues de laboratoires de recherche en sécurité IA ou de sociétés de cybersécurité ayant développé des compétences spécifiques sur les modèles de langage. La constitution d'un vivier d'évaluateurs qualifiés est un enjeu stratégique pour l'AI Office dans les prochains mois, plusieurs modèles systémiques majeurs devant être évalués simultanément.

Benchmarks et métriques d'évaluation GPAI recommandés

La documentation technique exigée par l'AI Act suppose de mesurer les performances du modèle GPAI sur des benchmarks reconnus. En 2026, les benchmarks de référence pour les obligations GPAI incluent plusieurs standards incontournables. Pour les **capacités générales**, le MMLU

(Massive Multitask Language Understanding) évalue la connaissance sur 57 domaines académiques ; le BIG-Bench Hard teste le raisonnement sur des tâches difficiles. Pour les **capacités de code**, HumanEval et MBPP restent les références. Pour l'**alignement et la sécurité**, TruthfulQA mesure la tendance aux hallucinations factuelles ; WMDP (Weapons of Mass Destruction Proxy) évalue les capacités dans les domaines CBRN sensibles — un benchmark particulièrement pertinent pour l'évaluation des risques systémiques. Pour les **capacités multilingues**, MGSM et FLORES-200 couvrent les performances en français et dans d'autres langues européennes. L'AI Office a exprimé sa préférence pour des évaluations reproductibles, documentées avec les hyperparamètres exact (température, top-p, nombre de shots). L'utilisation d'un framework standardisé comme **LM-Evaluation-Harness** d'EleutherAI est fortement recommandée pour garantir la reproductibilité.

Consommation énergétique des GPAI : l'obligation de reporting carbone

L'une des obligations souvent méconnues des modèles GPAI systémiques est le **rapport annuel de consommation énergétique** à transmettre à l'AI Office. Cette obligation reflète la prise de conscience croissante de l'impact environnemental des grands modèles de langage. En pratique, le rapport doit documenter la consommation énergétique totale de l'entraînement initial, des fine-tunings successifs et de l'inférence en production. Pour l'inférence, la mesure est complexe car elle dépend du volume de requêtes et de la configuration matérielle. Les métriques recommandées incluent la consommation en kWh par million de tokens générés, la **source d'énergie** (mix renouvelable vs fossile), et l'empreinte carbone équivalente en CO_2. Des frameworks comme **CodeCarbon** ou **ML CO2 Impact** permettent de mesurer automatiquement ces métriques durant l'entraînement. Pour les fournisseurs hébergeant leurs modèles dans des datacenters tiers (AWS, Azure, GCP), l'obtention des données de consommation réelle est soumise à la transparence des hyperscalers — un point de friction pratique largement documenté dans les retours de la consultation AI Office. La Commission a signalé son intention d'aligner progressivement ces obligations avec la directive sur le reporting de durabilité des entreprises (CSRD), créant un écosystème cohérent de mesure et de reporting de l'impact IA sur l'environnement.

Modèles GPAI et responsabilité civile : les nouvelles règles européennes

Parallèlement à l'AI Act, la directive européenne sur la **responsabilité du fait des produits défectueux (PLD révisée)** et la directive sur la responsabilité en matière d'IA (AI Liability Directive) restructurent le cadre de responsabilité civile applicable aux fournisseurs GPAI. La PLD révisée, entrée en application en 2026, étend le champ des "produits" aux logiciels — incluant les modèles GPAI — et introduit un mécanisme de **présomption de causalité** facilitant les recours des victimes. La AI Liability Directive crée un droit à la divulgation forcée de la documentation technique d'un système IA dans le cadre d'un litige, renforçant encore l'importance de la documentation GPAI. En pratique, les fournisseurs de modèles GPAI doivent anticiper que leur documentation technique, rédigée pour l'AI Office, pourra également être produite en justice. La rédaction doit donc être précise mais ne pas créer d'aveux implicites sur des capacités dangereuses qui ne seraient pas correctement gérées. L'articulation entre les obligations de transparence de l'AI Act et la protection du secret industriel dans les litiges civils est une question juridique non encore complètement résolue, qui fera probablement l'objet de jurisprudence dans les prochaines années. Les assureurs spécialisés commencent d'ailleurs à développer des produits couvrant spécifiquement le risque de responsabilité civile lié aux modèles GPAI, une tendance de fond qui devrait s'accélérer avec les premiers grands verdicts en matière de responsabilité IA.

Gestion des modèles GPAI dans la supply chain logicielle

L'intégration de modèles GPAI dans des produits logiciels crée une nouvelle dimension de la **supply chain security** que les RSSI doivent anticiper. Un modèle GPAI intégré via API ou déployé on-premise représente un composant tiers dont les propriétés — capabilities, biais, risques — doivent être auditées au même titre qu'une bibliothèque open source. L'AI Act renforce cette obligation en exigeant que les opérateurs connaissent les caractéristiques des modèles qu'ils utilisent. En pratique, les équipes DevSecOps doivent intégrer une étape de **vérification de conformité GPAI** dans leur pipeline CI/CD : vérification que le fournisseur a publié une model card conforme, que les limitations documentées sont compatibles avec l'usage

prévu, et que les conditions d'utilisation (AUP) n'excluent pas le cas d'usage envisagé. Des outils comme **MLflow Model Registry** ou **DVC** permettent de versionner les modèles et de tracer les changements de conformité. Un **AI Bill of Materials (AI-BOM)** — analogue au SBOM pour les logiciels — est une pratique émergente qui devrait devenir obligatoire pour les systèmes IA à haut risque.

Préparation aux contrôles de l'AI Office : ce qu'il faut anticiper

L'AI Office a annoncé le lancement de ses premiers contrôles de conformité GPAI pour le second semestre 2026. Ces contrôles peuvent prendre plusieurs formes. Les **contrôles documentaires** consistent en une demande de transmission de la documentation technique, du résumé des données et de la politique droits d'auteur — un délai de 30 jours est accordé pour la transmission. Les **enquêtes techniques** impliquent un accès au modèle pour des tests de capabilities et de sécurité conduits par l'AI Office ou des évaluateurs mandatés. Les **audits sur site** concernent principalement les modèles systémiques et permettent à l'AI Office d'accéder aux systèmes de formation du modèle et aux données d'entraînement. Pour se préparer, les fournisseurs doivent désigner un **point de contact unique** pour l'AI Office, maintenir un dossier de conformité à jour et consultable rapidement, et avoir défini en interne le périmètre des informations communicables (protection du secret des affaires via la procédure de l'article 78 du règlement). La désignation d'un représentant légal en Europe pour les fournisseurs établis hors UE est obligatoire dès août 2026.

Intersections avec le Digital Markets Act et le Cyber Resilience Act

Les fournisseurs GPAI opérant à grande échelle doivent également naviguer dans un écosystème réglementaire plus large que le seul AI Act. Le **Digital Markets Act (DMA)** peut s'appliquer si le fournisseur GPAI est désigné "gatekeeper" — ce qui concerne actuellement principalement les hyperscalers américains (Google, Microsoft, Apple). Le **Cyber Resilience Act (CRA)**, entré en

application en 2027, imposera des exigences de cybersécurité sur les produits numériques contenant des composants IA, créant une intersection avec les obligations de cybersécurité des modèles systémiques. Le **Data Act** régule le partage des données générées par les systèmes IA, avec des implications pour les fournisseurs qui collectent des données via leurs APIs. En France, la **loi SREN** (Sécurisation et Régulation de l'Espace Numérique) ajoute des couches de conformité spécifiques pour les services numériques dont les plateformes utilisant des GPAI. La cartographie de ces interactions réglementaires est un prérequis indispensable à toute stratégie de conformité cohérente.

Construire une culture de conformité IA dans l'entreprise

La conformité technique aux obligations GPAI ne suffit pas si elle n'est pas ancrée dans une culture organisationnelle adaptée. Les fournisseurs qui réussissent leur mise en conformité en 2026 sont ceux qui ont investi dans la **formation de leurs équipes techniques** aux principes de l'AI Act, au-delà des seules équipes juridiques et conformité. Les ingénieurs ML doivent comprendre pourquoi les décisions de benchmark, de sélection des données et d'architecture ont des implications réglementaires directes. Les **product managers** doivent intégrer les contraintes GPAI dès la conception des nouvelles fonctionnalités, évitant les remaniements coûteux en phase de lancement. La mise en place d'un comité **AI Ethics & Compliance** réunissant des représentants techniques, juridiques, métier et RH est une pratique de plus en plus répandue dans les entreprises françaises du secteur. Ce comité doit avoir un mandat clair et un accès direct au comité exécutif, car certaines décisions de conformité GPAI — notamment autour des données d'entraînement ou des évaluations de risques systémiques — ont des implications stratégiques et financières significatives.

Pour approfondir votre conformité AI Act, consultez notre guide général de conformité AI Act 2026, notre article sur les systèmes IA et leur classification, ainsi que notre analyse des systèmes agentiques sous l'AI Act. Pour les questions de protection des données liées aux LLM, notre guide chatbot entreprise et RGPD traite les points de friction pratiques. N'hésitez pas à consulter l'AI Office européen et les recommandations IA de la CNIL pour les dernières mises à jour réglementaires.