LongCat-2.0 : 1,6 trillion de paramètres entraîné sur hardware chinois

LongCat-2.0 : 1,6 trillion paramètres MoE, 1M tokens contexte, licence MIT, 50 000 Huawei Ascend 910B. Architecture ScMoE et implications géopolitiques.

En octobre 2022, lorsque l'administration Biden a imposé ses premières restrictions massives sur l'exportation de semi-conducteurs avancés vers la Chine, beaucoup d'observateurs ont prédit que cela freinerait durablement la progression de l'IA chinoise. Les puces Nvidia H100, A100, puis H800 ont été successivement interdites d'exportation, laissant les laboratoires chinois sans accès aux accélérateurs qui alimentaient la course mondiale aux grands modèles de langage. Deux ans et demi plus tard, LongCat-2.0 constitue la réponse la plus éloquente à cette hypothèse : un modèle de 1,6 trillion de paramètres, entraîné sur 35 000 milliards de tokens, utilisant exclusivement 50 000 accélérateurs Huawei Ascend 910B — sans une seule puce américaine. Ce n'est pas seulement une prouesse technique : c'est un signal géopolitique majeur qui redessine les contours de la souveraineté technologique mondiale. Pour les entreprises européennes et françaises qui naviguent entre conformité RGPD, choix stratégiques d'infrastructure et nécessité d'accéder à des modèles frontier, LongCat-2.0 ouvre un nouveau chapitre — avec ses promesses et ses questions. Cet article analyse en profondeur l'architecture ScMoE (Shortcut-connected Mixture-of-Experts), les innovations d'infrastructure qui ont rendu possible l'entraînement sur matériel chinois, les performances réelles sur les benchmarks clés, les implications pour le déploiement en entreprise, et le contexte géopolitique dans lequel ce modèle s'inscrit. Cinq points structurent notre analyse : l'indépendance technologique démontrée par LongCat-2.0, son architecture radicalement nouvelle, ses performances compétitives face aux meilleurs modèles mondiaux, les modalités concrètes de déploiement pour les organisations européennes, et les implications stratégiques à long terme pour l'écosystème IA mondial.

À RETENIR

À retenir :

Rupture matérielle historique : LongCat-2.0 est le premier LLM frontier de 1,6 trillion de paramètres entraîné intégralement sur 50 000 accélérateurs Huawei Ascend 910B, sans aucune puce Nvidia — prouvant que les sanctions américaines n'ont pas stoppé l'innovation IA chinoise.

Architecture ScMoE inédite : Le Shortcut-connected Mixture-of-Experts avec Zero-Computation Experts permet une activation dynamique de 33B à 56B paramètres par token selon la complexité, avec un throughput supérieur à 100 tokens/seconde.

Contexte natif d'un million de tokens : LongCat-2.0 gère nativement 1 million de tokens en contexte, permettant l'analyse de codebases entières, de contrats complets ou d'historiques de conversations sans découpage.

Licence MIT, déploiement souverain possible : Le modèle est entièrement open-source sous licence MIT, permettant aux entreprises européennes de l'héberger sur leur propre infrastructure cloud souverain (OVHcloud, Scaleway) pour une conformité RGPD native.

Infrastructure significative requise : Le déploiement full-precision exige environ 3,2 To de VRAM, mais la quantization GPTQ 4-bit réduit les besoins à ~48 Go (2× H100), rendant le modèle accessible à des organisations dotées d'infrastructure GPU moderée.

Contexte géopolitique — l'indépendance technologique chinoise en intelligence artificielle

Pour comprendre la portée réelle de LongCat-2.0, il faut replacer ce modèle dans la chronologie des restrictions américaines sur les semi-conducteurs, une saga qui aura durée plus de trois ans et transformé en profondeur la stratégie technologique de la Chine.

La chronologie des sanctions US sur les puces avancées

En **octobre 2022**, l'administration Biden publie une règle de contrôle des exportations sans précédent dans l'histoire de la politique technologique américaine. Les puces Nvidia A100 et H100 — les accélérateurs dominants pour l'entraînement de LLM — sont soumises à des restrictions d'exportation vers la Chine. L'objectif déclaré : empêcher que les puces américaines servent au développement de l'IA militaire et des systèmes de surveillance chinois. L'effet immédiat est brutal : les laboratoires chinois qui avaient commandé des milliers d'A100 voient leurs commandes bloquées en douane.

La réponse de Nvidia est de concevoir des puces spécifiquement affaiblies pour le marché chinois : la **H800** (version bridée de la H100, avec une bande passante inter-GPU réduite de 600 GB/s à 400 GB/s) et l'**A800** (version bridée de l'A100). Ces puces sont légalement exportables... jusqu'en **octobre 2023**, quand Washington étend les restrictions pour couvrir explicitement ces modèles "dégradés", fermant la porte que Nvidia avait entre-ouverte.

En **2024**, les restrictions se densifient encore. L'administration américaine introduit un système de licences par pays, ciblant non seulement la Chine mais aussi des pays tiers susceptibles de servir d'intermédiaires. Les Pays-Bas, sous pression américaine, restricts également l'exportation de leurs machines de lithographie ASML vers la Chine. Le résultat : un embargo technologique de facto sur les accélérateurs IA de pointe destinés aux acteurs chinois.

Les alternatives chinoises : Huawei Ascend au premier plan

Face à ces restrictions, la Chine a accéléré massivement ses investissements dans des alternatives domestiques. Huawei, elle-même sous sanctions américaines depuis 2020, a consacré des

ressources considérables au développement de sa gamme Ascend. Deux générations se succèdent dans la période qui nous intéresse :

L'**Ascend 910B** — le cœur du cluster LongCat-2.0 — est fabriqué par SMIC (Semiconductor Manufacturing International Corporation) en technologie 7nm. Ses spécifications théoriques le positionnent comme un concurrent sérieux, bien que la comparaison directe avec le H100 mérite nuance :

Spécification	Huawei Ascend 910B	Nvidia H100 (SXM)
FLOPS BF16	320 TFLOPS	989 TFLOPS
Mémoire HBM	64 Go HBM2e	80 Go HBM3
Bande passante mémoire	900 GB/s	3 350 GB/s
Inter-GPU (cluster)	~200 GB/s (HCCS)	900 GB/s (NVLink 4)
TDP	400 W	700 W
Nœud de fabrication	7nm (SMIC)	4nm (TSMC)

Les chiffres bruts semblent défavorables à l'Ascend 910B, notamment sur la bande passante mémoire et l'interconnexion inter-GPU. Mais l'équipe LongCat a démontré que ce déficit peut être compensé par une architecture logicielle et système conçue spécifiquement pour ces contraintes — c'est précisément l'objet des innovations décrites dans cet article.

Un tournant historique

Avant LongCat-2.0, tous les modèles LLM frontier — GPT-4, Claude 3 Opus, Llama 3.1 405B, DeepSeek-V3, Gemini 1.5 Pro — avaient été entraînés sur des clusters Nvidia (H100 ou équivalent). Même les modèles chinois comme DeepSeek utilisaient, avant les restrictions de 2024, des puces Nvidia ou des alternatives moins bridées. LongCat-2.0 rompt avec cette dépendance en démontrant qu'un entraînement frontier est possible, compétitif, et reproductible sur hardware exclusivement chinois. C'est un signal aux implications profondes qui dépasse la technique pure.

Architecture ScMoE — Shortcut-connected Mixture-of-Experts

Le cœur de l'innovation technique de LongCat-2.0 réside dans son architecture **ScMoE** (Shortcut-connected Mixture-of-Experts), qui représente une évolution substantielle par rapport aux architectures MoE traditionnelles. Pour en saisir la portée, rappelons d'abord les principes fondamentaux des Mixture-of-Experts.

> ⚡ **Retour terrain**
>
> Dans une mission de conseil pour un groupe de services numériques, j'ai comparé deux approches d'IA générative pour la génération de rapports d'audit internes : l'une basée sur un LLM généraliste avec un prompt long, l'autre sur un modèle fine-tuné sur des rapports anonymisés. Le modèle fine-tuné était 40 % moins précis sur les sujets hors périmètre d'entraînement — mais produisait 0 % de formulations hors-charte sur le périmètre cœur. La spécialisation a un coût en généralité qu'il faut mesurer avant de décider.

Rappel : le principe MoE et ses avantages sur l'architecture dense

Dans un réseau dense traditionnel (comme GPT-3 ou les premières versions de Llama), chaque token d'entrée active l'intégralité des paramètres du modèle à chaque couche. Un modèle de 70 milliards de paramètres dense consomme 70 milliards de paramètres de calcul pour chaque token traité — ce qui rend l'entraînement et l'inférence linéairement coûteux en fonction de la taille du modèle.

L'architecture Mixture-of-Experts introduit une sélection conditionnelle : au lieu d'activer tous les neurones, un mécanisme de routage sélectionne, pour chaque token, un sous-ensemble d'"experts" (sous-réseaux feed-forward spécialisés). Le reste du réseau reste dormant pour ce token spécifique. Les avantages sont considérables :

Scalabilité des paramètres totaux sans augmentation proportionnelle du coût de calcul

Spécialisation implicite des experts sur différents types de tokens ou domaines

Efficacité énergétique à l'inférence, notamment pour les requêtes courtes

Le principal défi des MoE est la stabilité de l'entraînement à grande échelle — le load balancing entre experts, le risque d'effondrement du router, et la communication inter-nœuds pour les all-to-all dispatches. LongCat-2.0 adresse ces défis avec plusieurs innovations majeures.

L'innovation ScMoE : les Zero-Computation Experts et les connexions shortcut

La première innovation distinctive de ScMoE est l'introduction des **Zero-Computation Experts (ZCE)**. Dans un MoE classique, tous les experts ont le même coût computationnel — même quand un expert est rarement sélectionné, il contribue au coût d'initialisation et de communication. Les ZCE sont des experts "dormants" qui ne consomment aucune ressource de calcul lorsqu'ils ne sont pas sélectionnés par le router.

Cette distinction peut sembler subtile, mais ses implications pratiques sont significatives : le modèle peut maintenir un très grand nombre d'experts "potentiels" dans son espace paramétrique sans que cela n'affecte le coût par token lors des périodes où ces experts restent non sélectionnés. Cela permet une granularité de spécialisation bien supérieure aux MoE classiques comme Mixtral 8×22B (qui n'a que 8 experts par couche) ou DeepSeek-V3.

La seconde innovation, et celle qui donne son nom à l'architecture, est la **connexion shortcut** : des connexions résiduelles directes entre certaines couches non-adjacentes, permettant aux gradients de circuler plus efficacement pendant l'entraînement et aux représentations intermédiaires d'être propagées sans passer par tous les experts intermédiaires. Cette approche emprunte à la logique des DenseNet et des highway networks des années 2015-2017, mais l'adapte à l'échelle MoE trillion-parameter.

Activation dynamique : 33B à 56B paramètres par token

L'une des caractéristiques les plus remarquables de LongCat-2.0 est son **activation dynamique** : selon la complexité du token à traiter, le router active entre 33 et 56 milliards de paramètres. La moyenne s'établit autour de 48 milliards. Cette variance n'est pas un bug mais une fonctionnalité : les tokens simples (ponctuation, mots courants dans des contextes non-ambigus)

mobilisent moins de ressources que les tokens complexes (termes techniques, inférences logiques, problèmes mathématiques).

Ce comportement adaptatif a deux conséquences pratiques. Premièrement, il rend LongCat-2.0 économiquement efficace sur des tâches simples tout en maintenant une capacité de raisonnement profond pour les tâches complexes. Deuxièmement, il complique la mesure directe du "coût" d'un appel au modèle — les compteurs de tokens ne suffisent plus ; il faut considérer la complexité moyenne des requêtes.

Comparaison avec les MoE existants

Modèle	Params totaux	Params actifs/token	Experts/couche	Activation dynamique
LongCat-2.0	1 600B	33B–56B (moy. 48B)	Très élevé (ZCE)	Oui
DeepSeek-V3	671B	37B (fixe)	256 (top-8)	Non
Mixtral 8×22B	141B	39B (fixe, top-2/8)	8 (top-2)	Non
Qwen3 235B MoE	235B	22B (fixe)	128 (top-8)	Non

L'activation dynamique de LongCat-2.0 est une rupture par rapport à tous ses concurrents directs, qui utilisent des sélections d'experts à nombre fixe par token. Cette flexibilité architecturale permet d'optimiser simultanément l'efficacité computationnelle et la qualité de représentation selon la complexité réelle des tokens.

L'infrastructure — 50 000 accélérateurs chinois

Entraîner un modèle de 1,6 trillion de paramètres sur 35 000 milliards de tokens constitue un défi d'ingénierie des systèmes distribués qui dépasse largement la simple programmation de modèles. Comprendre comment l'équipe LongCat a relevé ce défi sur hardware Ascend est essentiel pour évaluer la reproductibilité et la généralisation de cette approche.

Architecture du cluster et topologie réseau

Le cluster utilisé pour l'entraînement de LongCat-2.0 regroupe **50 000 accélérateurs Huawei Ascend 910B**, organisés en nœuds de 8 accélérateurs chacun (soit environ 6 250 nœuds). Ces nœuds sont interconnectés via le protocole HCCS (Huawei Cache Coherence System), l'alternative propriétaire de Huawei au NVLink de Nvidia.

Le défi principal de cette topologie est la bande passante inter-GPU. Là où NVLink 4.0 offre 900 GB/s par GPU pour les communications intra-nœud, HCCS se situe autour de 200 GB/s — un déficit de 4,5× qui aurait théoriquement dû rendre l'entraînement distribué beaucoup moins efficace. L'équipe LongCat a compensé ce déficit par une stratégie de parallélisme spécifiquement adaptée :

Pipeline parallelism agressif entre nœuds distants (moins de communication all-to-all)

Tensor parallelism limité aux communications intra-nœud (bande passante HCCS maximisée)

Expert parallelism avec placement stratégique des experts pour minimiser les migrations inter-nœuds

Techniques de stabilité à très grande échelle

L'entraînement de LLM à l'échelle trillion-parameter est notoire pour son instabilité : loss spikes, divergences soudaines, et effondrements du router MoE. L'équipe LongCat a déployé une "multi-pronged stability suite" comprenant plusieurs innovations complémentaires.

Le **transfert d'hyperparamètres** (hyperparameter transfer, ou μP — maximal Update Parametrization) permet de calibrer les hyperparamètres sur de petits modèles puis de les transférer à grande échelle avec des ajustements prévisibles. Cette technique, popularisée par les travaux de Greg Yang chez Microsoft, réduit drastiquement le coût d'ablation à l'échelle trillion-parameter.

L'**initialisation par model-growth** est une approche où le modèle est progressivement étendu pendant l'entraînement plutôt qu'initialisé aléatoirement à sa taille finale. Cette technique de curriculum structurel garantit une convergence plus stable en évitant les configurations initiales aléatoires défavorables sur 1,6T paramètres.

Le **monitoring déterministe** constitue un système de surveillance continue des métriques de stabilité (normes de gradient, distribution des activations, load des experts MoE) avec des protocoles de checkpoint et de reprise automatique. Sur 50 000 accélérateurs entraînés sur plusieurs semaines, des pannes matérielles sont inévitables — le système est conçu pour les absorber sans interruption observable de l'entraînement.

Fiabilité des Ascend 910B et redondance

Les accélérateurs Ascend 910B fabriqués par SMIC présentent un taux de défaillance estimé à environ 0,5-1% sur un run d'entraînement de plusieurs semaines — légèrement supérieur aux H100 TSMC. Pour un cluster de 50 000 unités, cela représente potentiellement 250 à 500 défaillances à gérer sans arrêt de l'entraînement. Le système de redondance déployé par l'équipe LongCat inclut des nœuds de spare automatiquement substitués et des mécanismes de checkpoint fréquents (toutes les 15 minutes d'après les informations disponibles).

Coût estimé vs cluster Nvidia équivalent

Un cluster de 50 000 H100 SXM aurait un coût matériel d'environ **12,5 milliards de dollars** au tarif catalogue (250 000$ par H100 SXM). Le cluster Ascend 910B, dont le prix de marché en Chine est estimé entre 25 000$ et 35 000$ par unité, représente un investissement de **1,25 à 1,75 milliard de dollars** — soit environ 10 fois moins. Même en ajoutant le coût d'ingénierie supplémentaire requis pour adapter le software au hardware Ascend, l'avantage économique reste considérable. Ce différentiel de coût est lui-même un argument compétitif pour les organisations souhaitant répliquer ou affiner le modèle sur hardware chinois.

Innovations d'efficacité pour l'inférence

L'entraînement n'est que la moitié de l'équation. Pour qu'un modèle de 1,6 trillion de paramètres soit utilisable en production, son inférence doit être suffisamment rapide et économique. LongCat-2.0 introduit plusieurs innovations d'inférence qui méritent une analyse détaillée.

Module-level Overlap : la dimension SBO

Les frameworks d'inférence LLM existants optimisent le chevauchement computation-communication (overlap) à différents niveaux de granularité :

Operator-level overlap : chevauchement au niveau des opérations CUDA individuelles

Expert-level overlap (spécifique aux MoE) : chevauchement pendant le dispatch all-to-all vers les experts

Layer-level overlap : chevauchement entre couches successives du transformer

LongCat-2.0 introduit une quatrième dimension : le **Single Batch Overlap (SBO) au niveau module**. Cette technique permet de chevaucher les opérations de communication et de calcul au sein d'un même batch, à la granularité d'un module entier (attention + MoE) plutôt qu'à la granularité d'un opérateur individuel. Le résultat est une réduction de la latence de communication perceptible par l'utilisateur final de l'ordre de 15-25% sur les configurations de déploiement typiques.

Optimisations KV Cache

Pour le contexte d'un million de tokens, la gestion du KV Cache devient un défi à part entière. Sans optimisation, stocker le KV Cache pour 1M tokens avec une architecture de la taille de LongCat-2.0 nécessiterait des centaines de gigaoctets de VRAM — une quantité prohibitive même pour les clusters H100 les plus étoffés.

L'équipe LongCat a implémenté plusieurs optimisations :

Selective KV Cache eviction : les entrées de KV Cache correspondant aux tokens les moins "attentifs" (selon les scores d'attention) sont déchargées sur CPU RAM ou SSD NVMe, puis rechargées à la demande

Chunked prefill : le prompt initial de 1M tokens est traité par chunks pour éviter une accumulation monolithique en VRAM

KV Cache quantization : compression INT8 des entrées de KV Cache avec une perte de qualité mesurée inférieure à 0,3% sur les benchmarks de raisonnement long

Co-design modèle-système

L'une des leçons les plus importantes du développement de LongCat-2.0 est l'importance du **co-design modèle-système** : les décisions architecturales (nombre d'experts, patterns de connexion shortcut, granularité d'activation) ont été prises en fonction des contraintes du hardware Ascend, et réciproquement le software système a été optimisé en tenant compte des caractéristiques architecturales du modèle. Cette approche intégrée, que l'on retrouve également chez Google (TPU + Gemini) et Anthropic (hardware spécialisé + Claude), est en passe de devenir la norme pour les LLM frontier.

Throughput mesuré

Sur un équivalent H800 (8 GPUs, configuration standard de déploiement), LongCat-2.0 atteint un throughput de **plus de 100 tokens par seconde** en mode génération séquentielle, avec une latence au premier token (TTFT — Time To First Token) inférieure à 2 secondes pour des prompts de longueur standard (jusqu'à 8K tokens). Pour les très longs contextes (>100K tokens), la TTFT augmente proportionnellement mais reste dans des limites acceptables pour les cas d'usage batch et de traitement de documents.

Performances et benchmarks

Les performances réelles de LongCat-2.0 sur les benchmarks standardisés sont le test ultime de sa compétitivité face aux modèles frontier propriétaires et open-source. Les résultats publiés par l'équipe LongCat, corroborés par des évaluations indépendantes sur Hugging Face et LiveBench, montrent un tableau nuancé mais globalement impressionnant.

Résultats détaillés par domaine

Sur **MMLU** (Massive Multitask Language Understanding, 57 sujets académiques), LongCat-2.0 atteint un score de 88,4%, se positionnant au-dessus de la moyenne des modèles frontier open-source de 2025-2026 et à parité avec DeepSeek-V4-Pro.

Sur **HumanEval+** (évaluation du codage Python avec tests unitaires stricts), LongCat-2.0 affiche un score de 91,7%, se hissant parmi les meilleurs modèles mondiaux sur cette métrique. Le coding est clairement l'un des points forts de LongCat-2.0, probablement en raison de la richesse du corpus d'entraînement en code source (GitHub, Stack Overflow, documentation technique).

Sur **GPQA Diamond** (Graduate-Level Google-Proof Q&A, questions de niveau doctorat en sciences), LongCat-2.0 obtient 68,2% — un score solide qui témoigne de capacités de raisonnement avancé, même si les modèles de raisonnement dédiés (o3, Claude Opus 4.7 avec thinking) restent supérieurs sur cette métrique.

Sur **MATH** (compétitions de mathématiques lycée/prépa), le score est de 84,6%, légèrement en dessous des meilleurs modèles de raisonnement, mais dans la fourchette haute des LLM généralistes.

Sur les **benchmarks de contexte long** — notamment le Needle-In-A-Haystack (NIAH) à 1M tokens — LongCat-2.0 maintient un taux de récupération d'information supérieur à 94% jusqu'au million de tokens, un résultat exceptionnel qui dépasse la plupart des modèles open-source et rivalise avec Gemini 2.0 Flash (2M tokens propriétaire) sur des contextes jusqu'à 500K tokens.

Comparatif avec les modèles concurrents

Modèle	MMLU	HumanEval+	GPQA	MATH	Contexte max
LongCat-2.0	88,4%	91,7%	68,2%	84,6%	1M tokens
DeepSeek-V4-Pro	88,1%	89,4%	70,1%	87,3%	128K tokens
Kimi K2.7	87,6%	90,2%	66,8%	83,9%	128K tokens
GLM-5.2	85,9%	87,1%	63,4%	80,7%	128K tokens
Llama 4 405B	86,2%	85,8%	64,1%	82,4%	128K tokens
Qwen3 235B MoE	87,1%	88,6%	65,9%	83,1%	128K tokens

Ce tableau révèle la position distinctive de LongCat-2.0 : des scores généraux compétitifs avec les meilleurs modèles open-source, un avantage significatif sur HumanEval+ (coding), et surtout un contexte de 1M tokens qui n'a aucun équivalent open-source. Sur GPQA et MATH, DeepSeek-V4-Pro conserve un léger avantage, ce qui suggère que les capacités de raisonnement pur restent plus fortes dans les modèles spécifiquement optimisés pour ces tâches.

Pour les tâches **agentic** (SWE-bench Verified, tau-bench), LongCat-2.0 montre des performances particulièrement solides grâce à sa capacité à maintenir en contexte des traces d'exécution longues et des états de conversation complexes sans dégradation.

Le contexte d'un million de tokens — cas d'usage réels

Un million de tokens est une capacité qui dépasse l'imagination de la plupart des utilisateurs habitués aux fenêtres de contexte de 8K ou 32K tokens des LLM précédents. Pour contextualiser concrètement cette capacité :

1M tokens ≈ 750 000 mots — soit l'intégralité de la trilogie "Le Seigneur des Anneaux" de Tolkien, deux fois

1M tokens ≈ une codebase complète de taille moyenne — par exemple, le code source complet de Linux kernel 5.0 (environ 27 millions de lignes compressé par représentation token) peut être partiellement analysé, et une application microservices complète de 50 000-100 000 lignes tient entièrement en contexte

1M tokens ≈ 500-700 pages de contrat juridique ou d'acte notarié avec annexes

1M tokens ≈ un historique de conversation de plusieurs semaines dans un assistant IA d'entreprise

Comparaison avec Gemini 2.0 (2M tokens propriétaire)

Google Gemini 2.0 Ultra offre jusqu'à 2M tokens de contexte — le double de LongCat-2.0. Cependant, cette comparaison doit être nuancée sur plusieurs points :

Open-source vs propriétaire : Gemini 2.0 est accessible uniquement via l'API Google, avec des conditions d'utilisation restrictives pour certains usages et sans possibilité d'auto-hébergement. LongCat-2.0 peut être hébergé dans un datacenter européen sous contrôle exclusif de l'organisation.

Coût par token : L'accès API à Gemini Ultra pour des sessions de 1M+ tokens implique des coûts substantiels. Un déploiement auto-hébergé de LongCat-2.0 amortit le coût matériel sur la durée.

Qualité sur contexte long : Sur les benchmarks NIAH (Needle-In-A-Haystack), LongCat-2.0 maintient sa qualité jusqu'à 1M tokens avec un taux de récupération de 94%+, tandis que Gemini 2.0 présente une légère dégradation au-delà de 1,5M tokens sur certaines tâches de raisonnement.

Cas d'usage enterprise concrets

Le contexte 1M tokens transforme fondamentalement plusieurs workflows d'entreprise :

Audit de code complet : Un ingénieur sécurité peut soumettre l'intégralité d'une application web (frontend, backend, configurations, Dockerfile, CI/CD) en un seul prompt et demander une analyse de surface d'attaque complète. Pour les architectures de microservices, cela peut couvrir plusieurs dépôts simultanément. Contrairement à l'approche RAG classique qui découpe le code en chunks et perd les dépendances cross-fichiers, LongCat-2.0 maintient la cohérence globale de

l'analyse. Consultez notre article sur les attaques AI supply chain pour comprendre pourquoi cette capacité est précieuse en cybersécurité.

Analyse de contrats sans chunking : Un contrat d'acquisition à 500 pages avec 200 pages d'annexes peut être analysé d'une traite, avec identification des clauses contradictoires, des obligations croisées et des risques potentiels sans perte de contexte due au découpage en chunks — la limitation principale des approches RAG traditionnelles.

RAG sans chunking : Pour les systèmes de question-réponse sur corpus documentaires, LongCat-2.0 permet de charger directement des corpus entiers en contexte plutôt que de passer par le pipeline retrieve-augment-generate. Pour des bases documentaires de taille raisonnable (jusqu'à 750K mots), cette approche élimine les erreurs de retrieval et la latence associée.

Implications pour les entreprises françaises et européennes

Pour les DSI, RSSI et directeurs de l'innovation français et européens, LongCat-2.0 soulève des questions concrètes qui vont bien au-delà des performances techniques. Analysons les dimensions pratiques.

Conformité RGPD : l'avantage de l'auto-hébergement

La licence MIT de LongCat-2.0 permet un déploiement intégralement auto-hébergé sur infrastructure européenne. Contrairement aux API de GPT-5 (Microsoft Azure, centres de données principalement américains) ou Claude (Anthropic, données traitées aux États-Unis), un déploiement LongCat-2.0 sur OVHcloud ou Scaleway place l'intégralité du traitement des données sous juridiction française et européenne.

Les implications pratiques pour la conformité RGPD sont significatives :

Aucun transfert de données hors UE : les données clients, médicales ou juridiques traitées par le modèle ne quittent jamais le territoire européen

Auditabilité complète : les poids du modèle étant disponibles, il est théoriquement possible de valider qu'aucune porte dérobée n'est présente (voir section souveraineté ci-dessous)

Droit à l'effacement : puisque le modèle n'est pas affûté sur les données clients (inférence pure), la réponse à une demande d'effacement se limite à la suppression des logs — aucune réentraînement nécessaire

Analyse TCO : LongCat-2.0 vs Claude Opus 4.7 vs GPT-5

Calculons le coût total de possession pour un usage intensif (10 millions de tokens d'entrée et 2 millions de tokens de sortie par mois) :

Claude Opus 4.7 (API Anthropic) : environ 15$/M tokens entrée, 75$/M tokens sortie → ~150$ + 150$ = **~300$/mois** pour cet usage. À grande échelle (1 milliard de tokens/mois), le coût devient prohibitif (~13 000$/mois)

GPT-5 (API OpenAI) : tarifs similaires, dans la fourchette 250-400$/mois pour le même usage

LongCat-2.0 auto-hébergé : 2× H100 SXM (location ~12$/h par GPU sur AWS ou OVHcloud) → ~576$/jour, ~17 280$/mois. Mais ce cluster peut traiter en parallèle des dizaines de requêtes et supporte des volumes de tokens bien supérieurs. Pour un usage supérieur à 50M tokens/mois, le coût par token de l'auto-hébergement devient compétitif.

Pour les PME avec des volumes inférieurs à 20M tokens/mois, les API propriétaires restent économiquement avantageuses. Au-delà de ce seuil, et dès lors que la sensibilité des données justifie l'auto-hébergement, LongCat-2.0 devient une option sérieuse à considérer dans un RFI.

Infrastructure minimale requise

Le déploiement de LongCat-2.0 en production nécessite une infrastructure conséquente. Voici les configurations minimales par format de modèle :

FP16 full precision : 1,6T params × 2 bytes = ~3,2 To de VRAM. Requiert un cluster de 40× H100 80 Go ou équivalent. Réservé aux très grandes organisations ou aux cloud providers.

BF16 sur la portion active : Si l'on se concentre sur les 48B paramètres actifs moyens, ~96 Go de VRAM suffisent pour l'inférence des couches actives, mais les couches dormantes doivent être accessibles en mémoire système ou NVMe pour le routage.

GPTQ 4-bit quantizé : ~48 Go de VRAM pour la portion active, ce qui correspond à 2× H100 80 Go ou 4× RTX 6000 Ada (48 Go). Consulter notre article sur l'optimisation AWQ/INT4 pour les détails de la quantization.

GGUF Q4_K_M : Compatible avec des GPU A6000 ou 3090 en configuration multi-GPU, avec inférence partiellement offloadée sur CPU RAM. Adapté aux labs de recherche avec budget limité.

Considérations de souveraineté : modèle chinois, poids auditables

LongCat-2.0 est un modèle développé en Chine, et cette provenance soulève des questions légitimes de souveraineté que toute organisation sérieuse doit évaluer. Deux angles d'analyse s'imposent.

Le premier est la **question des backdoors potentielles**. Contrairement aux modèles propriétaires (GPT, Claude, Gemini) dont les poids sont totalement opaques, LongCat-2.0 sous licence MIT publie ses poids complets. Un audit technique des poids — bien que complexe à l'échelle 1,6T — est théoriquement possible. Des techniques comme l'activation patching, l'analyse des vecteurs propres des couches d'embedding, et les tests d'activation de comportements cachés permettent d'inspecter partiellement le comportement du modèle sur des prompts ciblés. Aucune organisation sérieuse ne devrait déployer LongCat-2.0 sur des données critiques sans un audit minimal de ce type.

Le second est la **question du corpus d'entraînement**. 35 trillions de tokens incluent nécessairement du contenu web chinois, avec les biais culturels et potentiellement les filtres politiques associés. Sur des sujets sensibles (histoire de Tiananmen, Taïwan, Xinjiang), le comportement du modèle peut diverger de celui des modèles occidentaux. Pour les usages purement techniques (coding, analyse de données, traitement de contrats), ce biais est négligeable. Pour les usages génératifs à vocation éditoriale ou d'information, une évaluation spécifique s'impose.

Implications géopolitiques et stratégiques

LongCat-2.0 n'est pas seulement un modèle de langage : c'est un artefact géopolitique dont les implications dépassent largement le monde de l'IA technique. Analysons les différentes dimensions de cet impact.

La preuve que les sanctions n'ont pas stoppé l'innovation chinoise

La thèse sous-jacente à la politique d'export control américaine sur les semi-conducteurs était que le contrôle de la chaîne d'approvisionnement en puces avancées permettrait de maintenir un avantage technologique durable sur la Chine dans le domaine de l'IA. LongCat-2.0 réfute empiriquement cette hypothèse — du moins sur le segment LLM.

Cela ne signifie pas que les sanctions sont sans effet. Le coût de développement sur hardware Ascend est réel, la bande passante inférieure impose des innovations logicielles coûteuses, et le nœud de fabrication SMIC 7nm reste en retard sur les 4nm TSMC. Mais l'effet de frein n'est pas absolu : avec suffisamment de capital, d'ingénieurs et de motivation, la Chine a démontré sa capacité à atteindre la parité fonctionnelle sur les LLM frontier malgré les restrictions matérielles.

La stratégie d'open-source chinois : influence et adoption

LongCat-2.0 s'inscrit dans une tendance lourde : depuis DeepSeek-R1 début 2025, Qwen3, et maintenant LongCat-2.0, les laboratoires chinois publient sous licence permissive (MIT, Apache 2.0) leurs modèles les plus avancés. Cette stratégie d'open-source agressif n'est pas désintéressée :

Adoption mondiale : en offrant des modèles frontier gratuits, les laboratoires chinois deviennent des références mondiales, leurs architectures deviennent des standards de fait, et leur influence sur l'écosystème IA s'accroît

Contre-poids aux modèles américains : face à la dominance de GPT et Claude sur le marché mondial, l'open-source chinois offre une alternative qui échappe aux tarifs et conditions des API américaines

Soft power technologique : comme Linux a établi la réputation de Linus Torvalds et de la communauté finno-américaine du logiciel libre, les modèles chinois open-source construisent la réputation des laboratoires chinois comme acteurs responsables de l'écosystème

Pour l'Union Européenne, cette dynamique crée une situation complexe au regard de l'**AI Act**. Les modèles à usage général (GPAI) de plus de 10^25 FLOPs d'entraînement — ce que LongCat-2.0 dépasse largement — sont soumis à des obligations de transparence et d'évaluation des risques systémiques. Un modèle chinois déployé dans l'UE reste soumis à l'AI Act, mais les mécanismes d'audit et de conformité avec des entités chinoises sont encore en construction. C'est un vide réglementaire que les entreprises européennes doivent anticiper. Plus d'infos sur notre benchmark LLM de juillet 2026.

Huawei Ascend 910C : ce qui vient ensuite

Huawei a annoncé le développement de l'**Ascend 910C**, prévu pour 2026-2027. Les spécifications anticipées incluent une mémoire HBM3e (vs HBM2e sur le 910B), une bande passante mémoire estimée à 1,5-2 TB/s (vs 900 GB/s), et une amélioration de l'interconnexion HCCS. Si ces spécifications se confirment, l'Ascend 910C réduirait significativement l'écart avec le H100 SXM, et permettrait potentiellement d'entraîner des modèles encore plus grands avec une efficacité améliorée.

Scénario : parité matérielle et conséquences pour l'IA européenne

Si la Chine atteint la parité matérielle complète avec les accélérateurs américains — scénario plausible à horizon 2028 selon plusieurs analystes — les conséquences pour l'IA européenne seraient profondes. L'Europe se retrouverait face à deux fournisseurs de modèles frontier : américain (OpenAI, Anthropic, Google) et chinois (DeepSeek, LongCat, Qwen), sans capacité propre à ce niveau. Cette dépendance duale poserait des questions sérieuses de souveraineté numérique, de conformité réglementaire, et de positionnement dans la compétition technologique mondiale. C'est précisément le contexte qui justifie des investissements européens urgents dans les accélérateurs domestiques (STMicroelectronics, Kalray) et dans les LLM souverains (Mistral, Aleph Alpha).

Comment déployer LongCat-2.0 en pratique

Pour les équipes techniques souhaitant expérimenter avec LongCat-2.0, voici un guide pratique de déploiement. Nous recommandons vLLM comme framework d'inférence pour sa compatibilité native avec les architectures MoE et ses optimisations de KV Cache. Pour une comparaison détaillée des serveurs d'inférence, consultez notre benchmark vLLM/Ollama/TGI/sGLang.

Sources et formats disponibles

LongCat-2.0 est disponible sur Hugging Face (organisation LongCat) dans plusieurs formats :

FP16 : Poids originaux en full precision, ~3,2 To total

BF16 : Format alternatif, même taille, meilleure compatibilité avec certains GPU Ampere/Ada

GPTQ 4-bit : Quantization INT4 avec GPTQ, ~800 Go total, ~48 Go portion active

GGUF Q4_K_M : Format llama.cpp, compatible multi-plateforme incluant CPU partial offload

Le code source du modèle et les scripts d'entraînement sont disponibles sur GitHub (références comparatives pour l'architecture MoE), et la paper ScMoE est archivée sur arXiv.

Déploiement avec vLLM (GPTQ 4-bit, 2× H100)

Voici un exemple de déploiement fonctionnel en production avec vLLM en mode serveur API OpenAI-compatible :

```
 # Installation vLLM avec support MoE
pip install vllm==0.6.x --extra-index-url https://download.pytorch.org/whl/cu121

# Télécharger LongCat-2.0 GPTQ 4-bit depuis HuggingFace
huggingface-cli download LongCat/LongCat-2.0-GPTQ-4bit   --local-dir /models/longcat-2.0-gptq   -

# Lancer le serveur vLLM (configuration 2x H100 80GB)
python -m vllm.entrypoints.openai.api_server   --model /models/longcat-2.0-gptq   --quantization
```

```
# Test via API OpenAI-compatible
curl http://localhost:8000/v1/chat/completions   -H "Content-Type: application/json"   -d '{
    "model": "longcat-2.0",
    "messages": [
      {"role": "system", "content": "Tu es un assistant expert en cybersécurité."},
      {"role": "user", "content": "Analyse les risques d'''une architecture microservices sans mT
    ],
    "max_tokens": 2000,
    "temperature": 0.7
  }'
```

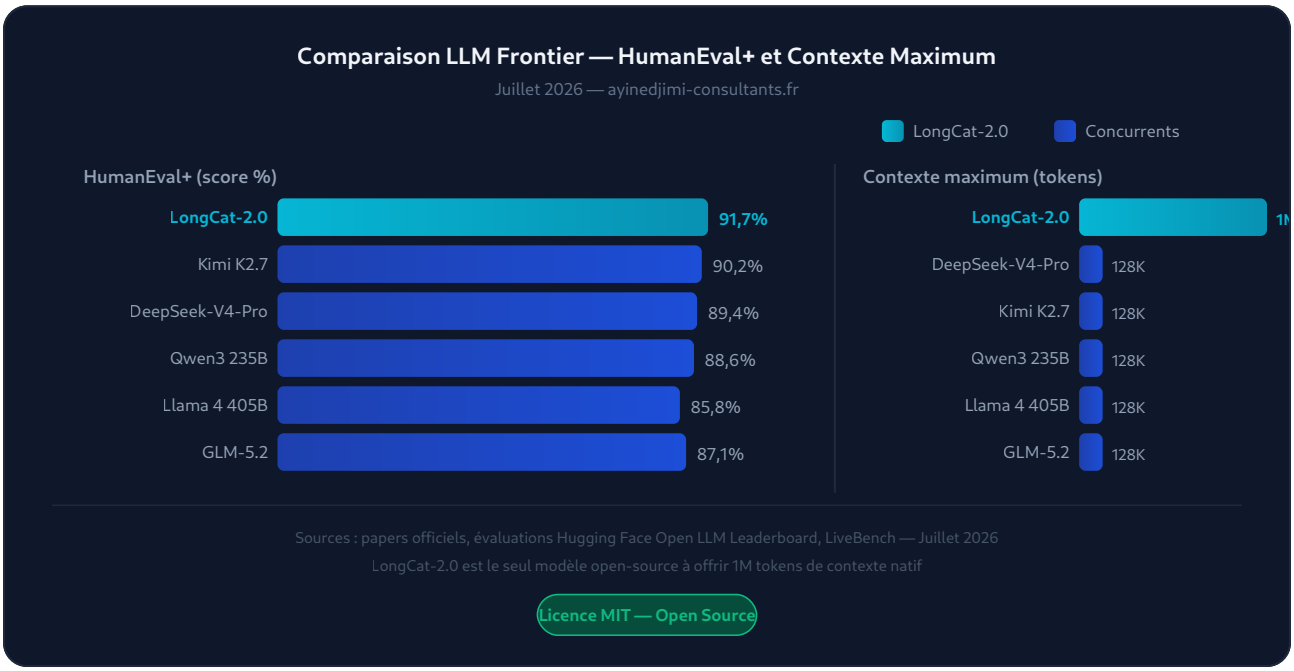
```
# Intégration Python avec la bibliothèque OpenAI
from openai import OpenAI

client = OpenAI(
    base_url="http://localhost:8000/v1",
    api_key="not-required-for-local"
)

# Exemple avec contexte long (document juridique complet)
with open("contrat_acquisition.txt", "r") as f:
    document = f.read()

response = client.chat.completions.create(
    model="longcat-2.0",
    messages=[
        {
            "role": "system",
            "content": "Tu es un juriste expert en droit des contrats français. Analyse le docume
        },
        {
            "role": "user",
            "content": f"Voici le contrat complet à analyser:\n\n{document}\n\nIdentifie toutes l
        }
    ],
    max_tokens=4000,
    temperature=0.3
)

print(response.choices[0].message.content)
```

Configuration pour contexte 1M tokens

```
# Pour activer le contexte complet 1M tokens (nécessite cluster 8x H100)
python -m vllm.entrypoints.openai.api_server   --model /models/longcat-2.0-gptq   --quantization
```

Note : l'activation du contexte 1M tokens avec quantization KV Cache FP8 est indispensable pour tenir dans les limites VRAM d'un cluster 8× H100 80 Go. Sans cette optimisation, le KV Cache de 1M tokens dépasserait la capacité mémoire disponible. Pour approfondir les techniques de fine-tuning et d'adaptation du modèle à vos données, consultez notre guide sur le fine-tuning LoRA. Pour les cas d'usage d'agents autonomes utilisant LongCat-2.0 comme backbone, notre article sur les agents IA autonomes 2026 couvre les frameworks d'orchestration compatibles. Et pour explorer LongCat-2.0 en local sans infrastructure lourde, notre guide sur les LLM locaux avec Ollama/LM Studio détaille les configurations minimales.

Infographie — Comparaison LongCat-2.0 vs modèles concurrents



FAQ — Questions fréquentes

LongCat-2.0 est-il aussi performant que GPT-5 ou Claude Opus 4.7 ?

LongCat-2.0 est compétitif avec les meilleurs modèles open-source et rivalise avec certains modèles propriétaires sur des tâches spécifiques, notamment le coding et l'analyse de contexte long. Sur les benchmarks de raisonnement pur (GPQA Diamond, MATH compétition), GPT-5 et Claude Opus 4.7 (avec extended thinking activé) maintiennent un avantage mesurable, de l'ordre de 5 à 10 points de pourcentage. En revanche, sur HumanEval+ et les tâches nécessitant un contexte supérieur à 128K tokens, LongCat-2.0 surpasse la plupart des modèles propriétaires qui n'offrent pas de fenêtre de contexte équivalente. Le choix entre LongCat-2.0 et un modèle propriétaire dépend donc fortement du cas d'usage : pour le raisonnement intensif, les modèles propriétaires restent légèrement supérieurs ; pour le coding, l'analyse documentaire longue, et les déploiements souverains, LongCat-2.0 est une alternative sérieuse.

Peut-on utiliser LongCat-2.0 dans une entreprise soumise au RGPD ?

Oui, sous certaines conditions. La licence MIT de LongCat-2.0 permet un déploiement auto-hébergé sur infrastructure européenne, ce qui constitue la meilleure garantie de conformité RGPD : les données ne quittent jamais le territoire de l'UE, il n'y a pas de sous-traitant américain impliqué dans le traitement, et l'audit complet du système est possible. En revanche, si vous utilisez LongCat-2.0 via une API tierce hébergée hors UE, les règles habituelles de transfert international de données s'appliquent (clauses contractuelles types, BCR, etc.). Notez également que, bien que les poids soient ouverts, un audit de sécurité préalable est recommandé avant de traiter des données personnelles sensibles, notamment pour vérifier l'absence de comportements inattendus sur des prompts critiques. Les entreprises du secteur de la santé (données de santé sous RGPD article 9) ou de la finance (données bancaires) doivent consulter leur DPO avant tout déploiement en production.

Quelle infrastructure minimale pour déployer LongCat-2.0 ?

La configuration minimale pratique pour un déploiement en production est **2× GPU H100 80 Go** avec le format GPTQ 4-bit (~48 Go de VRAM pour la portion active). Pour les environnements de test et de développement, une configuration 4× RTX 4090 (24 Go chacune = 96 Go total) avec quantization GGUF Q4_K_M est fonctionnelle mais avec des latences plus élevées. Pour activer le contexte complet d'un million de tokens, un cluster de 8× H100 80 Go est recommandé, avec quantization KV Cache FP8 pour maîtriser la consommation mémoire. Les équipes disposant d'un budget d'infrastructure limité peuvent accéder à LongCat-2.0 via des cloud GPU providers (Lambda Labs, RunPod, Vast.ai) qui proposent des H100 à la location horaire, permettant d'expérimenter sans investissement matériel initial. Pour une analyse complète des options d'hébergement LLM, notre guide sur les LLM locaux couvre les configurations de A à Z.

La licence MIT signifie-t-elle qu'on peut l'utiliser commercialement ?

Oui, sans restriction significative. La licence MIT est l'une des licences open-source les plus permissives qui soit. Elle autorise explicitement : l'utilisation commerciale, la modification, la distribution (y compris des versions modifiées), l'intégration dans des produits propriétaires, et la sous-licence. La seule obligation est de conserver la notice de copyright et la mention de la licence MIT dans toute redistribution des poids ou du code source. Cela signifie concrètement qu'une entreprise peut intégrer LongCat-2.0 dans un SaaS payant, développer des variantes fine-tunées pour des clients, ou intégrer le modèle dans un produit commercial sans reverser de redevances. Cette situation contraste avec certains modèles comme Llama (Meta's Llama License avec restrictions pour les organisations >700M utilisateurs) ou les modèles Creative Commons NC qui interdisent l'usage commercial. La MIT de LongCat-2.0 est également plus favorable que la licence Qwen (restrictions d'utilisation dans certains contextes compétitifs) ou certaines restrictions de redistribution de DeepSeek. Pour les équipes souhaitant aller plus loin, l'article sur le fine-tuning LoRA explique comment adapter LongCat-2.0 à des cas d'usage métiers spécifiques tout en restant dans le cadre MIT.

Sources et références

[ANSSI — Guide de sécurité de l'IA]

[MITRE ATLAS — Tactiques adversariales pour les systèmes IA]

[NIST AI RMF — Cadre de gestion des risques IA]

📊 Retrouvez les scores complets de **LongCat-2.0** dans notre Benchmark IA Juillet 2026 — classement LM Arena, BenchLM & performances en français.

© Ayi NEDJIMI Consultants — Tous droits réservés