La Guerre des SLMs (Small Language Models) & Edge AI

2 December 2025 • 3 min de lecture

LLM SLM Agentic

> *La course au gigantisme (plus de paramètres, plus de GPU, plus de bras et plus de chocolat...) ralentit au profit de modèles ultra-optimisés qui tournent à l'edge ou sur de petits devices. Si vous utilisez encore un modèle à 100 Mds de paramètres pour résumer un email en 2026, vous brûlez du cash. La maturité IA ne se mesure plus à la taille du modèle, mais à l'efficience de son architecture.*

Il y a un an, on pensait que la course à l'IA se résumait aux "gros" modèles (plus de paramètres, plus de GPU). Fin 2025, les "petits" modèles surfent sur la vague du ROI, pendant que la lame de fond des gros l'engloutit.

Pour Gartner, d'ici 2027, les modèles spécialisés seront trois fois plus utilisés en entreprise que les modèles généralistes : le budget alloué à la GenAI ne fond plus. Pour des tâches d'exécution routinières, un SLM réduit la facture d'inférence.

2025 a marqué le basculement : on fait désormais tourner des modèles performants directement sur les NPU des laptops et smartphones. C'est une réponse au cauchemar de la conformité (RGPD/AI Act). Avec une architecture Edge, la donnée sensible est traitée de manière locale. La surface d'attaque change, la latence est maîtrisée et le coût du cloud devient nul pour ces opérations (j'espère que M. OPEX et M. CAPEX ne m'en tiendront pas rigueur).

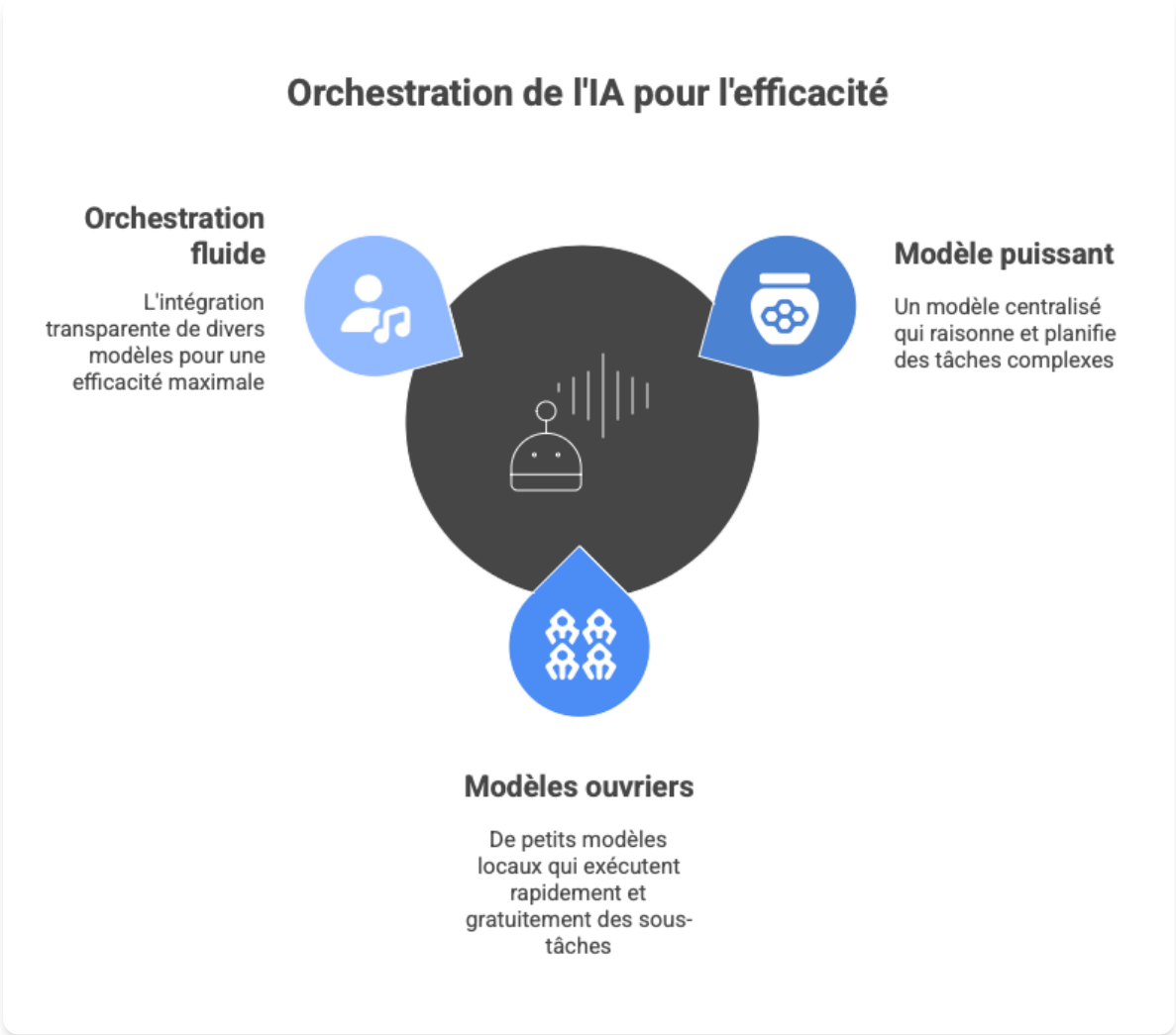
La stratégie pour 2026 n'est pas binaire (gros vs petits), elle est hiérarchique. L'erreur est de croire que le "cerveau" central doit être le modèle le plus lourd.

Les recherches récentes (notamment *ToolOrchestra*) expérimentent l'inverse :

- Un orchestrateur : un modèle compact, sur-entraîné spécifiquement pour l'usage d'outils, bat désormais de gros modèles (GPT-5) sur la planification et le routage de tâches complexes et spécifiques. Il ne "sait" pas tout, mais il sait parfaitement "qui" appeler. On peut même "durcir" l'orchestration avec des stratégies du type Guided LLM.
- Des exécutants : une équipe de modèles spécialisés ou d'API qui exécutent les tâches unitaires.

Forrester confirme cette vision : on passe du chatbot monolithique à l'*Agentic AI*. L'enjeu n'est plus la connaissance encyclopédique du modèle, mais sa capacité à manipuler le système d'information.

Une voie se renforce vers une décentralisation de l'intelligence; nous sommes maintenant dans un monde où modèles frugaux, modèles AGI de taille conséquente, modèles d'orchestration, modèles de raisonnement cohabitent et peuvent être agencés en fonction des problèmes pour optimiser la robustesse, le ROI, les aspects sécuritaires ou réglementaires.



Sources & Liens

- **[1] Gartner Prediction :** *Small AI Models to Outpace LLMs 3-to-1 by 2027* - Lire le rapport
- **[2] Forrester Tech Trends :** *Agentic AI in Top 10 of Emerging Technology for 2025* - Lire l'analyse
- **[3] Microsoft Research (Le précurseur) :** *Phi-3 Technical Report* - Papier Arxiv (Avril 2024)
- **[4] Arxiv (SOTA 2025)**: *ToolOrchestra: Elevating Intelligence via Efficient Model and Tool* - Arxiv (Nov 2025)
- **[5] Arxiv**: *Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation* Arxiv

Lego I sac 1

Les opinions exprimées dans cet article sont strictement personnelles et ne reflètent pas nécessairement celles de mon employeur. Les contenus sont fournis à titre informatif et ne constituent pas un conseil juridique. Cet article explore des concepts architecturaux émergents et analyse des tendances de marché (Gartner, Forrester). Les solutions technologiques citées le sont à titre d'exemple et ne préjugent pas des choix technologiques ou des partenariats de mon employeur

---

Eric Blaudez - AI Innovation Strategist & Technical Lead
La Guerre des SLMs (Small Language Models) & Edge AI

LinkedIn