L'IA générative dans l'administration française : le cas d'« Albert »

Par Ulrich TAN
Ingénieur en chef des Mines et chef du DataLab à la direction interministérielle du Numérique (DINUM)

Dans cet article, nous explorons le projet « Albert », l'intelligence artificielle générative développée par la direction interministérielle du Numérique (DINUM). La stratégie sous-jacente émane d'une vision technologique ambitieuse pour l'administration : s'approprier une technologie nouvelle et prometteuse, pour ne pas la subir, et assurer la souveraineté numérique de l'État. Lancé en réponse à l'essor fulgurant de l'IA, Albert vise à fournir des solutions adaptées aux besoins de l'administration française tout en garantissant la confidentialité des données et la puissance de calcul nécessaire. Nous discuterons des trois piliers du projet, ses applications concrètes, et les défis technologiques et éthiques à relever pour une utilisation optimale de l'IA dans le secteur public.

Dans cet article, nous nous intéressons à « Albert », présenté comme l'intelligence artificielle générative souveraine de l'État français. La stratégie sous-jacente émane d'une vision technologique ambitieuse pour l'administration : s'approprier une technologie nouvelle et prometteuse, pour ne pas la subir, et assurer la souveraineté numérique de l'État.

Genèse et contours du projet : maîtriser ses données en plein *boom* de l'IA

Le projet Albert débute concrètement en juin 2023[1], soit six mois après la sortie de ChatGPT (30 novembre 2022). Le succès sans précédent de l'application d'OpenAI[2] a accéléré l'engouement général autour de l'IA, et l'administration n'a pas échappé à ce mouvement. Les initiatives se multiplient alors dans la sphère publique pour explorer les promesses de l'IA et en particulier l'IA générative.

Dans ce contexte, Albert a été lancé par le DataLab de la DINUM pour répondre à deux limitations récurrentes : la confidentialité des données et la puissance de calcul. Lorsque vous utilisez un service comme ChatGPT, les données que vous lui envoyez sont en effet captées par les serveurs de l'entreprise américaine, dont vous utilisez également la puissance de calcul. À l'inverse, les solutions déployées par le DataLab dans le cadre d'Albert le sont sur des serveurs maîtrisés par l'administration, ce qui lui impose par là-même de disposer d'une puissance de calcul suffisante.

La technologie repose sur les grands modèles de langage (ou LLM pour *Large Language Models*) et leur intégration dans des systèmes d'IA (SIA) permettant de générer du texte, ou du code[3], à partir d'une instruction formulée en langage naturel (fournie dans une invite textuelle, un « prompt », ou à partir d'un média comme du son ou une image). Elle s'inscrit ainsi dans la catégorie des IA génératives, une famille d'IA qui faisait déjà parler d'elle avec les hypertrucages (*deepfakes*) quelques années avant l'émergence des LLM[4]. Ces SIA peuvent couvrir un large spectre d'usages : résumer du texte, générer des réponses, assister un agent dans sa recherche d'information, transcrire une visio-conférence ou une déposition, faire une traduction à la volée…

Il faut noter que la DINUM ne développe pas directement des LLM. Cela nécessite en effet des ressources dont elle ne dispose pas. Elle utilise donc des modèles préexistants (modèles pré-entrainés, ou modèles de fondation), qu'elle adapte (*fine tuning*, *prompt engineering*) et intègre dans des SIA (ajoutant des capacités comme la génération augmentée par récupération documentaire – RAG, *Retrieval Augmented Generation*). La DINUM a d'ailleurs été la première administration française civile à faire du *fine tuning* et à intégrer du RAG dès l'été 2023. La diversité des modèles disponibles, en particulier les modèles dont l'usage est libre comme

[1] Le projet n'a été nommé « Albert » que lorsqu'il a été dévoilé publiquement le 5 octobre 2023.

[2] L'application connait la plus forte croissance jamais enregistrée, atteignant 100 millions d'utilisateurs actifs par mois en moins de deux mois. Source : https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[3] La génération de son, d'image ou encore de vidéo est également possible, mais ne fait actuellement pas partie des cas d'usage étudiés.

[4] Un historique complet et rigoureux des IA génératives dépasse largement le cadre de cet article. On mentionnera simplement l'apparition des réseaux antagonistes génératifs en 2014 [1], et celle des modèles auto-attentifs, ou *tranformers*, en 2017 [2]. Ces derniers sont utilisés dans la majorité des LLM aujourd'hui (mais il existe d'autres architectures de LLM).

ceux de Mistral AI[5], permet à Albert de ne pas dépendre d'un seul fournisseur de modèle. Le DataLab a d'ailleurs déployé plus d'une dizaine de modèles différents et change régulièrement de modèles de fondation[6]. Tous ces déploiements sont faits dans le respect de ses normes de confidentialité et de sécurité[7].

Pas juste un modèle : Trois piliers pour s'approprier l'IA

Face à la diversité d'usages anticipée, Albert a été pensé comme un dispositif technologique. Albert n'est donc pas à proprement parler une IA (il n'y a pas un seul modèle ou un seul SIA), il recouvre en fait une gamme de services et de produits d'IA, bâtie autour de trois piliers :

- le sur-mesure, où le DataLab développe une solution reposant sur les LLM, en co-construction directe avec les utilisateurs finaux ;
- l'IA à la demande (IA *as a service*), où il s'agit de mettre à disposition des services prêts à l'emploi, consommables directement *via* des applications en ligne (API en ligne ou logiciel en tant que service – SaaS, *Software as a Service*) ;
- le commun numérique où les codes, les modèles, les jeux de données, mais aussi les résultats de recherches sont publiquement partagés de manière ouverte et sous licences libres, et peuvent faire l'objet de contributions externes, notamment de la communauté du logiciel libre (*open source*).

Premier pilier

Le premier pilier a été assez naturellement le premier à être actif. C'est celui qui permet non seulement de monter en compétence, mais aussi de favoriser l'adoption des outils en assurant qu'ils soient bien adaptés aux besoins des utilisateurs.

On peut par exemple citer le cas d'Albert France Services, développé avec l'Agence nationale de la cohésion des territoires (ANCT). Il s'agit de fournir aux agents des France Services un assistant virtuel avec lequel ils peuvent échanger pour chercher une solution à un problème concret d'un usager. Le projet a été lancé en conditions opérationnelles en janvier 2024, et couvre aujourd'hui 44 France Services dans 6 départements. L'enjeu est double : vérifier que la technologie est suffisamment pertinente, *i.e.* que l'IA répond bien aux questions posées, et s'assurer que la solution est utile aux agents. On notera que la diversité des situations rencontrées par France Services est un véritable défi, et il a fallu itérer à plusieurs reprises en composant avec les limites des modèles actuels[8].

Ces expériences sont précieuses pour l'administration, qui se forge ainsi un véritable savoir-faire dans la conduite de projet d'IA générative. Par ailleurs, les développements sur Albert France Services, en particulier le RAG, sont majoritairement repris dans tous les autres projets de la famille Albert (mutualisation et externalités positives).

Deuxième pilier

Le deuxième pilier (IA *as a service*) a ainsi largement bénéficié des travaux précédents. La première brique du pilier 2 a été le déploiement d'Albert API, qui met à disposition des administrations des fonctionnalités d'IA (notamment des LLM) via une API[9] en ligne. Albert API, en service depuis juin 2024, est le premier service en ligne de ce type déployé sur des serveurs SecNumCloud. Il permet aux administrations de faire l'économie d'un déploiement d'IA (mobilisant puissance de calcul et compétences) pour leurs projets de SIA utilisant des LLM[10]. Les premiers utilisateurs ont rapporté un gain de temps pouvant aller jusqu'à trois mois dans le développement de leurs SIA.

Aujourd'hui, 25 projets dans l'administration publique utilisent Albert API. Il a par exemple permis d'automatiser des tâches de résumé, qui prenaient deux à trois jours habituellement, et qui prennent désormais moins d'une journée à être traitées[11]. Albert API est utilisé par LaSuite (suite bureautique collaborative de la DINUM), qui a mis à disposition des capacités d'IA générative dans son éditeur de texte (Docs) et son tableur (Grist). Tous les agents utilisant ProConnect ont de fait accès à Albert[12].

Le pilier 2 est dès lors un vecteur structurant de la diffusion de l'IA générative dans l'administration, rendant l'IA accessible aux agents de l'État, directement dans leurs outils du quotidien (comme avec Docs). Il est encore en construction, et doit être complété par une offre de logiciels en tant que services, tels qu'un agent conversationnel généraliste[13].

[5] Mistral AI propose à la fois des modèles libres et ouverts (*i.e.* dont les poids sont accessibles), mais aussi des modèles propriétaires payants.

[6] Il utilise aussi bien des modèles de Mistral que ceux de Meta, Lighton, Google ou encore PleIAs.

[7] Le DataLab utilise en particulier des serveurs répondant à la certification de sécurité SecNumCloud, délivrée par l'Agence nationale de la Sécurité des systèmes d'information (ANSSI). Il échange régulièrement avec l'ANSSI sur les questions de sécurité, ainsi qu'avec la Commission nationale de l'informatique et des libertés (Cnil) sur le sujet de la protection des données personnelles.

[8] Au moins trois modèles spécialement ré-entrainés ont été testés, et l'interface utilisateur a régulièrement évolué en fonction des retours.

[9] Interface de programmation applicative (API, *Application Programming Interface*).

[10] Par ailleurs, Albert API respecte les normes imposées par les usages (celles d'OpenAI), de sorte qu'un projet développé sur une API LLM différente peut facilement migrer vers Albert API (*via* redirection d'URL).

[11] Dans le cas d'usage rapporté : la machine peut traiter en 20 minutes l'ensemble des textes à résumer, et les agents n'ont plus qu'à vérifier les résultats.

[12] LaSuite est accessible aux agents en se connectant simplement *via* ProConnect (https://www.proconnect.gouv.fr/). Aujourd'hui, 1 000 agents béta-testent l'IA dans Docs.

[13] Un premier agent conversationnel a été développé et testé à titre purement expérimental au DataLab dès le mois de novembre 2023. Un autre, accessible directement *via* la messagerie Tchap (Albert Tchap), a également été développé et présenté en juin 2024, puis ouvert à tous les agents de la DINUM en septembre 2024. Plusieurs versions ont été testées (dont une version multimodale). Ces expérimentations permettent de tester en conditions réelles les questions de protection des données personnelles, et de sécurité informatique.

Troisième pilier

Enfin, le troisième pilier est celui qui doit sceller l'appropriation collective de la technologie. Il vise à la mobilisation d'une communauté numérique où les administrations, avec la communauté *open source*, mais aussi les acteurs privés (éditeurs et / ou intégrateurs de logiciels, opérateurs, start-ups, hébergeurs *cloud*…) se seront réapproprié entièrement le projet[14].

Les codes, les modèles, et les jeux de données utilisés par Albert sont librement disponibles et ouverts à contributions sur les plateformes publiques de référence Github (https://github.com/etalab-ia/), et Hugging Face (https://huggingface.co/AgentPublic). Le DataLab y a notamment mis à disposition le plus grand jeu de données administratives ouvertes au monde[15]. L'intégralité des données ouvertes de la direction de l'Information légale et administrative (DILA), incluant le site Légifrance, y est en particulier disponible sous une forme directement interrogeable par les SIA (format « vectorisé »)[16, 17].

Ce troisième pilier se construit sur un temps long. Les interventions en conférences, formations, les publications ou encore l'animation de communautés participent à ce travail.

Les perspectives : faire émerger les applications à forte valeur ajoutée

Albert est donc avant tout un programme de mise à disposition technologique (expertise, services, infrastructure, commun numérique), répondant à une forte demande. Il permet en particulier de réduire le risque de fuite de données par une utilisation inappropriée de services numériques non encadrés (*shadow* IT). En complément d'Albert, la DINUM a mis en place un incubateur de produits IA, Alliance[18]. Ce dispositif a pour objectif premier d'orienter les projets IA de l'administration vers la réussite au sens de l'impact réel des solutions proposées. C'est un sujet que ne traite que partiellement Albert, *via* son pilier 1 (les projets du pilier 1 sont d'ailleurs menés dans le cadre d'Alliance). Dans une start-up, Albert serait le CTO[19], là où Alliance jouerait plutôt le rôle du CPO[20].

Par ailleurs, dans la perspective d'un déploiement durable de la technologie, les défis restent nombreux. La frugalité des SIA est un enjeu majeur. Dans ce domaine, le DataLab travaille sur la taille optimale des modèles dans un esprit *lean* (chercher les modèles les plus petits possibles à niveau de performance donné). Cela nécessite d'évaluer correctement les performances des IA, ce qui reste un domaine de R&D très actif[21].

Des limites inhérentes aux LLM posent également des questions de gouvernance. Ces IA font des erreurs, voire « hallucinent » (ce sont des modèles statistiques). C'est d'autant plus trompeur que, non seulement les réponses générées ont une forme assertive, mais aussi, plus elles seront performantes, plus elles seront crédibles. En cas d'erreur, cela pose des problèmes du fait de l'opposabilité des réponses d'une administration, soumise au Code des relations entre le public et l'administration (CRPA). Dans ces conditions, ces IA sont aujourd'hui destinées à être une aide aux agents, qui restent responsables de la relation avec les usagers. Mais les agents doivent être formés pour être conscients des limites des IA. De plus, l'explicabilité n'est pas assurée avec ces algorithmes, au sens des obligations de transparence algorithmique. Ces IA ne doivent donc pas être utilisées dans un processus de décision administrative.

Conclusion

Cet aperçu des défis à relever est loin d'être exhaustif. Plutôt que chercher à être holistique, l'approche du DataLab, et plus largement de la DINUM, se veut pragmatique avec des méthodes itératives (agile), et visant la frugalité (*lean*). La transformation numérique de l'État met à l'épreuve les capacités d'adaptation de l'administration. Dans le domaine de l'IA générative, l'État s'est montré pro-actif, et la France peut se féliciter de faire partie des pays les plus en avance sur le sujet.

Références

[1] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., BING X., WARDE-FARLEY D., OZAIR S., COURVILLE A. & BENGIO Y. (2014), "Generative adversarial nets", *Proceedings* of the 27th International Conference on Neural Information Processing Systems, MIT Press, 10 juin (arXiv 1406.2661).

[2] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017), "Attention is all you need", *Advances in Neural Information Processing Systems*, 30, 12 juin, p. 15 (arXiv 1706.03762).

[14] Le code étant *open source*, tout le monde peut le réutiliser librement.

[15] 380 milliards de tokens : https://huggingface.co/datasets/AgentPublic/open_government

[16] Pour faire des recherches documentaires, par exemple pour du RAG, les SIA en question ont besoin de vectoriser les données, c'est-à-dire les transformer en représentations numériques (on parle aussi de plongement sémantique ou *embedding*), https://huggingface.co/datasets/AgentPublic/DILA-Vectors

[17] Ce travail s'inscrit en fait dans le cadre d'un projet plus vaste, encore en développement (Albert Data).

[18] https://alliance.numerique.gouv.fr/

[19] *Chief Technology Officer*.

[20] *Chief Product Officer*.

[21] Le DataLab mettra prochainement à disposition un outil d'évaluation en ligne (projet EG1 en interne).