

# SpeedUP



## Study LLM4Gov

<b>Donneur d'ordre</b>	Ministère de la Digitalisation
<b>Objectifs de l'appel à solutions</b>	<p>Le ministère de la Digitalisation a reçu une multitude de demandes concernant la mise en place d'un « Large Language Model » (LLM) dédié aux besoins du secteur public. L'objectif principal de l'intégration d'un tel modèle est d'automatiser une variété de tâches. Ces tâches comprennent principalement la génération de contenu et de résumés. Le modèle devrait également être capable de générer des réponses cohérentes aux questions concernant les besoins métier spécifiques.</p> <p>Comme première étape avant la mise en œuvre potentielle d'un ou plusieurs LLM, une étude visant à identifier la meilleure solution est à réaliser dans le cadre de cet appel à solutions.</p>
<b>Type de solution recherchée</b>	Etude de faisabilité, benchmark de l'existant et proposition d'un concept
<b>Critères de sélection</b>	<ul style="list-style-type: none"><li>• Qualité de l'offre soumise (approche, structure, niveau de détail, exhaustivité) - 40%.</li><li>• Calendrier et proposition de prix - 30 %.</li><li>• Profils des consultants (expérience pertinente) - 30 %.</li></ul>

<b>Standards à respecter</b>	NA
<b>Détails concernant l'IP et autres</b>	<p>Les livrables sont couverts par les lois sur la propriété intellectuelle et les droits d'auteur et resteront la propriété exclusive du ministère de la Digitalisation pendant toute la durée du projet et au-delà.</p> <p>Tous les litiges concernant ce projet sont régis par le droit luxembourgeois et les tribunaux du Grand-Duché de Luxembourg sont exclusivement compétents pour recevoir et traiter ces litiges.</p>
<b>Délai du dépôt de l'offre</b>	19 novembre 2024
<b>Calendrier du suivi</b>	La durée du projet ne doit pas dépasser 6 mois. Le lancement est prévu pour janvier 2025 au plus tard.
<b>Contact en cas de question</b>	<a href="mailto:datascience@digital.etat.lu">datascience@digital.etat.lu</a>
<b>Délai pour les questions</b>	6 novembre 2024

**Descriptif du  
problème/défi à  
résoudre**

Une étude visant à identifier le(s) modèle(s) le(s) mieux adapté(s) aux besoins du gouvernement luxembourgeois est la première étape avant une éventuelle mise en place d'un ou plusieurs LLM. Afin d'identifier les besoins des organismes du secteur public concernant l'utilisation des « Large language models », une enquête a été réalisée pour recueillir leurs besoins. Les résultats de l'enquête seront communiqués à l'opérateur économique sélectionné. Les cas d'utilisation principaux sont l'analyse de documents, un outil de recherche alternatif ainsi que la création de contenu.

Outre les résultats de l'enquête, le partenaire économique sélectionné sera également informé sur plusieurs cas d'utilisation (certains mentionnés ci-dessous dans la section 2.1) d'une complexité similaire qui ont déjà été identifiés pour un modèle de LLM au sein du gouvernement.

Ce projet vise à produire une étude pour le ministère de la Digitalisation classant les LLM existants selon les critères pertinents ci-dessous, ainsi qu'un plan d'implémentation d'un ou de plusieurs outils LLM.

La liste suivante est une liste non exhaustive de critères qui doivent au minimum être abordés dans l'étude, et donc également inclus dans le rapport final :

**Sélection des modèles :**

- Les principaux LLM de référence doivent au minimum être pris en compte (à la fois propriétaires et à open-source).
- Les LLM avec différents nombres de paramètres doivent être pris en compte (du plus léger au plus sophistiqué).
- Les LLM utilisant différentes technologies doivent être pris en compte (par exemple transformateurs, decoder-only, encoder-only, mixture of experts).

**Capacités et caractéristiques :**

- Capacités générales : les capacités globales des LLM candidats doivent être fournies.
- Capacités spécifiques aux cas d'utilisation : la capacité de chaque LLM candidat à fournir une réponse pertinente pour chaque cas d'utilisation doit être évaluée.

**Versatilité :**

- La capacité de chaque LLM à couvrir plusieurs cas d'utilisation avec des solutions de haute qualité doit être présentée.
- La capacité de chaque LLM à intégrer facilement de nouveaux cas d'utilisation.  
De plus amples informations sur les cas d'utilisation possibles seront fournies ultérieurement (résultats de l'enquête).

**Support linguistique :**

- Il convient d'évaluer si le modèle doit prendre en charge l'anglais (EN), le français (FR), l'allemand (DE) et le luxembourgeois (LU).
- Les LLM doivent également être capables de comprendre et de générer du code dans différents langages de codage.

### **Analyse comparative :**

- Les capacités et la versatilité des différents LLM doivent être démontrées par des critères d'évaluation pertinents, tels que décrits ci-dessus.
- Les LLM doivent être évalués sur la base de différents cas d'utilisation, y compris le risque d'hallucination, la vitesse et la pertinence.
- Les caractéristiques de l'environnement de référence doivent être indiquées, ainsi que tous les paramètres pertinents pour garantir la reproductibilité des résultats obtenus.

### **L'intégration**

#### **Intégration avec les modèles existants :**

- Il convient d'analyser la faisabilité de l'intégration de modèles existants (comme les modèles de « Speech to text », un chatbot ciblant un domaine spécifique, etc.) De plus amples informations sur ces modèles seront fournies à l'opérateur économique choisi.
- Disponibilité de l'API : Le modèle doit fournir une API accessible et fiable pour l'intégration avec les outils existants.

#### **Hébergement et infrastructure :**

- Il convient d'évaluer les avantages et les inconvénients de l'hébergement en nuage par rapport à l'hébergement sur site, des solutions sous licence par rapport aux solutions sans licence, et des solutions à code source ouvert par rapport aux solutions propriétaires.
- La capacité à traiter des données sensibles (personnelles ou autres) dans un environnement souverain et sécurisé doit être évaluée.
- La capacité du LLM à être déployé sur une plateforme de conteneurs OpenShift doit être évaluée.
- La capacité pour le LLM d'être déployé sur des plates-formes air-gapped telles que des plates-formes cloud souveraines déconnectées doit être évaluée.

#### **L'intégration dans un environnement complet doit être évaluée. Cela inclut :**

- La facilité d'intégration avec des outils tels que, mais sans s'y limiter :
  - Les bases de données vectorielles
  - Les caches
  - Les outils de monitoring, de logging et de tracing
  - Les outils d'orchestration
  - Les agents
- L'API REST du modèle doit supporter OAuth2.0 et OIDC.
- La possibilité de mettre en œuvre les LLM sur différents environnements (*Build and run environment*).
- L'existence d'une suite complète intégrant le LLM.

### **Maintenance et intégration :**

Il convient de prendre en compte les versions, les mises à jour de modèles, la reproductibilité des réponses générées, les efforts de maintenance, les exigences matérielles (CPU, GPU, mémoire, stockage) et la consommation d'énergie.

- Cela doit être fait de manière générale et spécifiquement pour les cas d'utilisation prévus.

### **Coût**

#### **Analyse des coûts :**

- Une analyse complète des coûts doit être effectuée, y compris l'estimation du coût de l'hébergement, de la maintenance, des frais de licence potentiels ou des ressources humaines. Afin de comparer les LLM, les coûts doivent être indiqués en TCO (coût total de possession) pour une période de 5 ans.

D'autres critères pourraient découler des réponses à notre enquête qui sera communiquée lors du lancement du projet.

### **Conformité**

#### **Protection des données :**

- Le modèle doit être conforme au GDPR.
- L'exposition du LLM aux attaques d'extraction doit être fournie.

**Pour des informations complètes concernant les exigences, les livrables et le type d'offre à soumettre, veuillez-vous référer aux spécifications du projet sur le site web.**