AI READER CAM

Réponses aux questions reçues

QUESTION 1

Where do we submit the proposal?

Réponse / Answer

FR : L'offre doit être soumise via courriel à l'adresse cam@cam.etat.lu

EN : The offer must be sent via email to cam@cam.etat.lu

QUESTION 2

What are the requirements for the proposal ?

Réponse / Answer

FR : Les informations nécessaires mentionnées dans le document 'Cahier de charge AI Reader CAM' sous '3.3 Content of the offer', sont à soumettre en 1 document.

EN : The necessary information mentioned in the specifications document 'Cahier de charge AI Reader CAM' under '3.3 Content of the offer', should be submitted in 1 document.

QUESTION 3

Les agents des CAM préféreraient-ils un affichage visuel des zones détectées sur les documents, ou une simple extraction textuelle suffit-elle à leurs besoins ?

Réponse / Answer

FR : Pour les agents, une extraction textuelle est suffisante.

EN : For the end-user, a text output is sufficient.

QUESTION 4

Que voulez-vous dire par : "Ideally, it should be capable of self-training to identify label/value pairs in cases where it has already been trained to find values in multiple variations of the same type of document" ? (phrase présente dans la partie "Description of the problem / challenge" du document "speedup-ai-reader-en.pdf")

Réponse / Answer

FR : Nous sommes conscients qu'il existe des systèmes capables d'identifier eux-mêmes les zones de recherche après avoir été entraînés sur un certain nombre de variantes au préalable. Pour répondre à cet appel, vous ne devez pas nécessairement remplir cette condition et pouvez proposer une alternative.

EN : We are aware that there are systems capable of identifying search areas on their own after being trained on a certain number of variants beforehand. To answer to this call you don't necessarily need to fulfill this condition and can suggest an alternative.

QUESTION 5

You mention intellectual property in your call for tenders. Can you specify what this refers to? Which aspects of the source code or algorithms will be exclusively owned by CAM? Do you accept intellectual property-based approaches, tailored to your needs of course (meaning you only have a license to use the product but not on the product itself)?

Réponse / Answer

FR : La solution spécifique, le code, la configuration et les modèles produits pour l'AI Reader CAM doivent être la propriété intellectuelle du CAM / État luxembourgeois. Si un modèle de licensing, avec des frais recurrents s'appliquent, ces coûts sont à indiquer clairement avec une prévision annuelle.

EN : The specific solution, code, configuration and models produced for the AI Reader CAM becomes the intellectual property of the CAM / Luxembourg State. If a solution requiring recurrent licensing fees is chosen, those costs must be clearly laid out, including an annual prevision.

QUESTION 6

There is no relevant self-learning solution on the market. Do you accept solutions including a learning loop based on user feedbacks and data-scientists tuning?

Réponse / Answer
FR : Il n'y a aucun critère d'exclusion concernant les solutions à la problématique posée. La reconnaissance optique de caractères (OCR) par deep learning présente ses défis, et nous étudierons chaque proposition en fonction de son équilibre entre faisabilité réaliste et promesses potentiellement exagérées.

EN : There are no exclusion criteria regarding the solutions to the issue at hand. Deep learning OCR has its challenges, we will study every proposition for its balance between realistically feasible or slightly overpromising.

QUESTION 7

Is a remote access by our data scientists to the target environments possible (VPN, SFTP, ...)?

Réponse / Answer
FR : L'accès à distance est disponible via VPN (Cisco Secure Client et l'utilisation d'une carte à puce Luxtrust) après la signature d'un contrat, incluant un accord de confidentialité (NDA).

EN : Remote access is available via VPN (Cisco Secure Client and the usage of a Luxtrust Smartcard) after having signed a contract, including an NDA.

QUESTION 8

What are the expectations regarding the future extension of the solution to other use cases or document types?

Réponse / Answer
FR : La solution doit être neutre par rapport au type de document. Cela signifie que nous avons proposé 2 types de documents et leurs variantes pour ce projet. La solution doit être entraînable pour d'autres types de documents et leurs variantes.

EN : The solution should be document type neutral. It means that we proposed 2 types of documents and their variants for this phase. The solution should be trainable for other document types and their variants.

QUESTION 9

Should future integrations be partially developed in this POC or simply conceptually validated?

Réponse / Answer
FR : L'intégration future doit être documentée de manière à nous convaincre que ces extensions sont possibles sans avoir à réécrire une grande partie de la solution. Bien entendu, si un service est simulé et démontré de manière convaincante, cela ne nous pose aucun problème.

EN : The future integration should be documented such that we can be convinced that those extensions are feasible without rewriting a significant portion of the solution. Of course if a service is mocked up and demonstrated in a convincible matter, we don't mind.

QUESTION 10

Would you agree to follow-up during the first three months, then optimized quarterly versions afterwards?

Réponse / Answer
FR : Nous étudierons chaque offre ainsi que leurs coûts de fonctionnement et évaluerons les différentes approches qui pourraient émerger en termes de suivi.

EN : We will study each offer and their running costs and evaluate the different approaches which might emerge in terms of follow-up.

QUESTION 11

Should the points in paragraph 3.3 "Content of the Offer" be specified on the scope of the POC or the full underlying project scope?

Réponse / Answer
FR : Le projet lui-même est la preuve de concept (POC), qui nécessite de livrer un système capable d'extraire les données des 2 types de documents et de leurs variantes. Il n'y a pas d'autre projet sous-jacent, si ce n'est que nous attendons de la solution qu'elle soit entraînable pour d'autres types de documents et que nous décrivions les fonctionnalités (capacité API) qui seront nécessaires dans les phases suivantes.

EN : The project itself is the POC, which requires to deliver a system capable of extracting the data of the 2 types of documents and their variants. There is no other underlying project, except that we expect the solution to be trainable for other

document types and we describe features (API capability) which will be needed in further phases.

QUESTION 12

What are your criteria for evaluating the success of the data extraction engine (processing time, error rate, etc.)?

Réponse / Answer
FR : Les critères sont principalement axés sur l'expérience utilisateur, et la métrique principale est la facilité d'utilisation ainsi que la rapidité avec laquelle le résultat est disponible pour l'utilisateur. Nous évaluons également l'architecture globale et sa cohérence technique en termes de conception et de réutilisabilité. Il n'y a pas de valeurs précises à atteindre. Chaque soumission sera mise en concurrence avec les autres.

EN : The criteria are very user experience centric and the main metric is how easy it is to use and how quickly the output is available to the user. We also assess the overall architecture and its technical consistency in terms of design and reusabilty. There are no precise values to be met. Every submission will be put in competition the others.

QUESTION 13

What is the expected user experience for the user interface (simplicity, advanced options, etc.)?

Réponse / Answer
FR : L'aspect le plus important est de réduire le temps de traitement d'un dossier. Bien que cette phase ne gère pas l'ensemble du processus de transfert de données de machine à machine, nous souhaitons néanmoins limiter le nombre de clics au minimum.

Idéalement : les utilisateurs glissent et déposent les documents dans un dossier cible, puis retrouvent les données dans l'interface web. Ils peuvent ensuite copier/coller les informations dans l'application cible.

À l'avenir : le système opérationnel enverra les documents via une API au backend du moteur, qui renverra le texte extrait via API au système opérationnel. L'utilisateur pourra alors valider ou corriger les données extraites.

Dans les deux scénarios, une interface administrateur est nécessaire pour visualiser les extractions et gérer les actions proposées en matière d'auto-apprentissage ou de correction du modèle.

EN : The most important aspect is to reduce the time to process a case. Although this phase does not handle a full chain of machine to machine data transfer, we still want to do as few clicks as possible.

Ideally: Users drag and drop documents into a target folder, find the data in the web interface, then proceed to copy / paste to the target application.

Future: Operational system sends documents over an API to the engine's backend, the latter has the output text come back via an API to the operational system where the user then validates / corrects the input.

In both scenarios we need an admin interface to visualise the extraction and operate whatever has been proposed in terms of 'self learning' or corrective action to the model.

QUESTION 14

What are the CTIE environment hardware capabilities, specially regarding CPU, GPU and RAM (exact models)?

Réponse / Answer

FR : Le CTIE propose des machines virtuelles, dont certaines avec des capacités GPU (NVidia "Tesla V100" GPU), spécifiquement dédiées aux applications de machine learning. Les environnements sont évolutifs en termes de RAM et de vCPUs. Des informations détaillées sur le type ou les modèles de matériel ne sont pas disponibles pour le moment. Le système d'exploitation peut être Linux Ubuntu, Redhat ou Windows Server.

EN : The CTIE proposes virtual machines , some with GPU capabilities (NVidia "Tesla V100" GPU), specifically dedicated to ML applications. The environments are scalable in terms of RAM and vCPUs. Detailled information on hardware type or models is not available at this time. OS can be Linux Ubuntu or Redhat or Windows Server environment.

QUESTION 15

The project specification mentions that "The contract falls under the fixed global price regime." Should we include our proposed pricing in the submitted proposal, or is there a predetermined fixed budget assigned for the CAM project?

Réponse / Answer
FR : Vous devez soumettre un prix dans votre offre, sachant qu'il s'agit d'un projet à prix fixe. Définition : Un contrat à prix fixe est un contrat dans lequel le prix convenu pour le travail reste inchangé tout au long du projet. Peu importe si plus de temps, de matériaux ou de main-d'œuvre sont nécessaires que prévu initialement, le prix reste le même.

Le budget prévu pour ce projet est de 100.000 EUR HTVA.

EN : You should submit your price proposition for the offer, knowing that it is a fixed price project. Definition : A fixed-price contract is a contract where the agreed-upon price remains unchanged throughout the project. It doesn't matter if more time, materials or labor must be used than first estimated, the price remains unchanged.

The budget for this project is EUR 100,000 excluding VAT.

QUESTION 16

Although maintenance is not explicitly listed among the expected deliverables, would you require maintenance support for the installed system for a specific period?

Réponse / Answer
FR : Les frais d'opération et de maintenance sont à mentionner dans l'offre. Si ces coûts ne sont pas encore prévisible à ce stade, ils doivent être inclus comme livrable dans le projet.

EN : Running costs, if any, and maintenance costs should be mentioned in the offer. If these costs are unclear at this stage, they should be included as a deliverable in the project

QUESTION 17

Can you provide concrete examples of the (anonymised) documents for the initial phase (VISAs and seafarer's booklets)?

Réponse / Answer

FR : L'annexe avec les exemples illustratifs sur le site GovTechLab a été temporairement supprimée, mais une nouvelle version est à nouveau disponible :

https://govtechlab.public.lu/dam-assets/speedup/ai-reader-cam/data-to-be-extracted-from-documents.pdf

EN : The attachment with the illustrative samples on the GovTech Lab website was temporarily removed, but a new version is available again: https://govtechlab.public.lu/dam-assets/speedup/ai-reader-cam/data-to-be-extracted-from-documents.pdf

QUESTION 18

What are the expected performance indicators (e.g., extraction rate, acceptable error rate)?

Réponse / Answer

FR : Voir question 12.

EN : See question 12.

QUESTION 19

Are there specific language needs for documents (French, English, etc.) or are they mainly in French?

Réponse / Answer

FR : Les documents numérisés sont généralement en anglais. Si vous parlez de la documentation de la solution, elle peut être en anglais ou en français.

EN : The scanned documents are usually in English. If you are talking about the documentation of the solution, it can be in English or French.

QUESTION 20

Do you want the application to be containerized for easy deployment? Is a solution like Kubernetes/OpenShift being considered? and Do you already have an infrastructure with these solutions??

Réponse / Answer

FR : Bien que le CTIE dispose d'une infrastructure pour utiliser Kubernetes, l'utilisation et le déploiement d'une solution conteneurisée ne sont pas dans le

cadre de ce projet. Nous préférons une approche non conteneurisée pour le POC, ce qui n'exclut pas de passer aux conteneurs dans un projet ultérieur.

EN : Although the CTIE has an infrastructure to use Kubernetes, using and deploying a containerized solution is not in the scope of this project. We prefer an non containerized approach for the POC, which does not exclude to switch to containers in a later project.

QUESTION 21

What level of modularity do you expect for future integration with the GESTCAM app?

Réponse / Answer

FR : Nous attendons de la solution qu'elle dispose, pour les documents entrants, d'une implémentation de balayage de dossiers pour ce POC et d'une capacité API pour l'intégration avec le futur GESTCAM. Cette API devra être capable d'acquérir des documents et de renvoyer les résultats.

EN : We expect the solution to have, for the incoming documents, a folder scanning implementation for this POC and an API capability for the integration with the future GESTCAM. This API will have to be able to acquire documents and send back the results.

QUESTION 22

How will the training data (anonymised documents) be made available? Will there be continuous access during the project?

Réponse / Answer

FR : Pour le développement, un minimum de documents sera mis à disposition. Une fois qu'une version entièrement déployable sera disponible, elle sera installée sur l'environnement de test du CAM. Là, le système devra être entraîné avec toutes les données d'entraînement disponibles.

EN : For the development a minimum of documents will be made available. Once a full deployable version is available, it will be installed on the CAM's Test environment. There the system should get trained with all training data available.

Question 23

Are there any specific requirements related to the retention of data extraction and validation logs for traceability reasons?

Réponse / Answer

FR : Bien qu'il devrait y avoir des journaux d'erreurs et d'activités, nous n'avons pas besoin que les données extraites soient elles-mêmes conservées.

EN : While there should be error and activity logs, we don't need the extracted data itself to be preserved.

Question 24

What will be the extent of support expected for the configuration and installation by the CAM IT team?

Réponse / Answer

FR : L'équipe informatique du CAM peut déployer des solutions sur des environnements Linux et Windows, à condition qu'un manuel complet d'installation et de configuration soit disponible. Une assistance pour cette étape doit être prévue.

EN : The CAM IT can deploy solutions on linux and windows environments, provided a complete installation and configuration manual is available. Assistance for this step has to be accounted for.

Question 27

How will the process of validating the extracted data be managed? Will human operators need a structured review process and should the system provide feedback to improve accuracy?

Réponse / Answer

FR : Les opérateurs devraient avoir la possibilité, sur l'écran présentant le jeu de résultats, de déclencher une réponse de rétroaction, en fonction du niveau de profondeur que votre solution met en œuvre pour corriger le modèle.

EN : The operators should have a possibility, via the screen presenting the results, to trigger a feedback response, depending on the level of depth that your solution provides to correct the model.

Question 28

In what format and definition are the documents/images stored and provided? Is there a pre-treatment that is carried out in order to reframe them? or are they supplied in RAW?

Réponse / Answer

FR : Les documents sont fournis soit sous forme d'images (jpeg, png) soit sous forme d'images en PDF. Aucun prétraitement n'est effectué par les agents. Les utilisateurs déposent tel quel les documents entrants provenant de diverses sources dans le système.

EN : The documents are provided either as images (jpeg, png) or images in PDF. There is no pre-processing done by the agents. The users drop the incoming documents directly from the various sources into the system.

Question 29

What are the network, firewall or proxy constraints within the CTIE infrastructure to be taken into account for communication between the different modules (frontend, backend, databases, etc.)?

Réponse / Answer

FR : Idéalement, le POC devrait être déployable sur une seule instance. Comme la solution n'est pas exposée au public, il ne devrait y avoir aucune contrainte bloquante de la part de l'équipe Firewall pour permettre la communication entre le front-end et le back-end.

EN : Ideally, the POC should be deployable on one instance. As the solution is not exposed to the public, there shouldn't be a blocking contraint from the firewall team to allow communication between front and back.

Question 30

Can the documents contain official signatures or stamps that need to be detected and interpreted (e.g. date of signature, stamp of the issuing body)?

Réponse / Answer

FR : Aucun tampon ni symbole manuscrit ne devra être extrait.

EN : No stamps or handwritten symbols have to be extracted.

Question 31

What is the approximate volume of documents to be processed daily or monthly?

Réponse / Answer

FR : Approximativement 450 documents par mois.

EN : Approximately 450 documents per month.

Question 32

Do you have any constraints in terms of image size or resolution (e.g. maximum 20 MB per file, resolution of 300 dpi, etc.)?

Réponse / Answer

FR : Aucune contrainte. La plupart des fichiers font moins de 1 Mo. Plus de 5 Mo est extrêmement rare.

EN : No constraints. Most files are under 1MB. >5MB is extremly rare.