

Projet d'Initiation à la Recherche ou Projet d'Innovation-Recherche (barrer)

Nom du laboratoire ou de l'entreprise/établissement : CGI

TUTEUR(S)

NOM-Prénom Tel Mel

- PIERRE François-Xavier Tel 06 03 18 21 87 Mel francois-xavier.pierre@cgi.com

- LABETOULLE Thomas Mel thomas.labetoulle@cgi.com

Tuteurs INSA si projet industriel:

Prénom NOM

Mel:

TITRE DU PROJET

Coût unitaire de l'inférence LLM : banc de mesure et comparaison des stratégies d'optimisation

MOT-CLES

FinOps, inférence LLM, IA Generative, coût par token, unit economics, optimisation, quantization, caching, benchmark, tableau de bord

DESCRIPTIF (RESUME), indiquer l'enjeu sociétal de l'INSA de Toulouse s'il y a lieu

Contexte. La généralisation des applications fondées sur les grands modèles de langage déplace une part croissante des coûts informatiques vers l'inférence, dont le prix se mesure à la requête ou au token. Faute de mesure comparative rigoureuse, les organisations peinent à arbitrer entre les leviers d'optimisation disponibles.

La question de recherche est la suivante : ***à qualité de réponse constante, quel gain de coût unitaire chaque stratégie permet-elle réellement d'obtenir, et à quel prix en latence ?***

État de l'art (semestre 1). Économie unitaire de l'inférence LLM : décomposition du coût par token, comparaison entre un modèle publique exposé en API et un modèle auto-hébergé, puis revue des principaux leviers d'optimisation (batching, , routage de modèles selon la difficulté de la requête...).

La thématique peut être élargie au cadrage FinOps du coût de l'intelligence artificielle générative.

Réalisation (semestre 2). Conception d'un banc de mesure comparant le coût par requête de plusieurs stratégies sur une tâche fixée, avec restitution par tableau de bord. Chaque stratégie est évaluée sur trois axes : coût unitaire, latence et qualité de réponse mesurée par une métrique à définir.

Le résultat attendu prend la forme d'un enseignement chiffré : telle stratégie divise le coût par N à qualité comparable.

Livrables : article scientifique en anglais (état de l'art, méthode, résultats), banc de mesure reproductible et tableau de bord comparatif.

PROFIL DES ETUDIANTS SOUHAITE (1 seul choix par projet)

- ☐ AE-SE : spécialité Automatique-Electronique parcours Systèmes Embarqués
- ☒ IR-SI : spécialité Informatique parcours Systèmes Informatiques
- ☐ IR-SC : spécialité Informatique parcours Systèmes Communicants

- ☐ (optionnel) ce projet peut être proposé à un ou des étudiants d'échange sur la partie réalisation seule (semestre 1 et/ou semestre 2)

PRIORITE : si vous posez plusieurs sujets, indiquer ici la priorité de ce sujet (1= plus prioritaire) 1