LLM 4 Cloud

Évaluation des capacités des LLM dans la génération de fichiers de configuration pour le Cloud

Auteurs

Adam OUZEGDOUH
Pierre LAURENT

Encadrants

Hayet BRABRA
Walid GAALOUL

Partenaires

ensIIE
Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise

Site vitrine

Présentation

Évaluation de **5 LLM** sur la génération de fichiers Infrastructure as Code (IaC) Terraform pour AWS, à travers **458 problématiques** de différents niveaux de difficulté, **en français et en anglais.**

Basé sur le benchmark **IaC-Eval** de l'Université du Michigan et de Cisco Research

LLM testés :

CodeGemma	DeepSeek-Coder-V2	Phi-4	Qwen-2.5-Coder	Codestral
7B	16B	14B	14B	22B

Scénarios (jeux de données) :

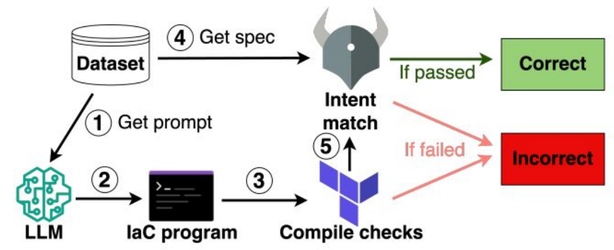
À partir de **458 problématiques** issues du dataset de Iac-Eval, réparties en 6 niveaux de difficulté, trois sous-ensembles de données ont été créés :

- Un premier sous-ensemble regroupant les problématiques de **niveaux de difficulté 1 à 3** (253 problématiques).
- Un deuxième sous-ensemble regroupant les problématiques **de niveaux de difficulté 4 à 6** (205 problématiques).
- Un troisième sous-ensemble regroupant les problématiques de **niveaux de difficulté 1 à 3, traduites en français** via le LLM Gemma 3 (12B).

Informations diverses :

- **Site vitrine** permettant de visualiser les résultats.
- Adaptation du benchmark pour utiliser des **LLM en local avec Ollama**.
- Conteneurisation du benchmark avec **Docker** pour une **meilleure portabilité**.

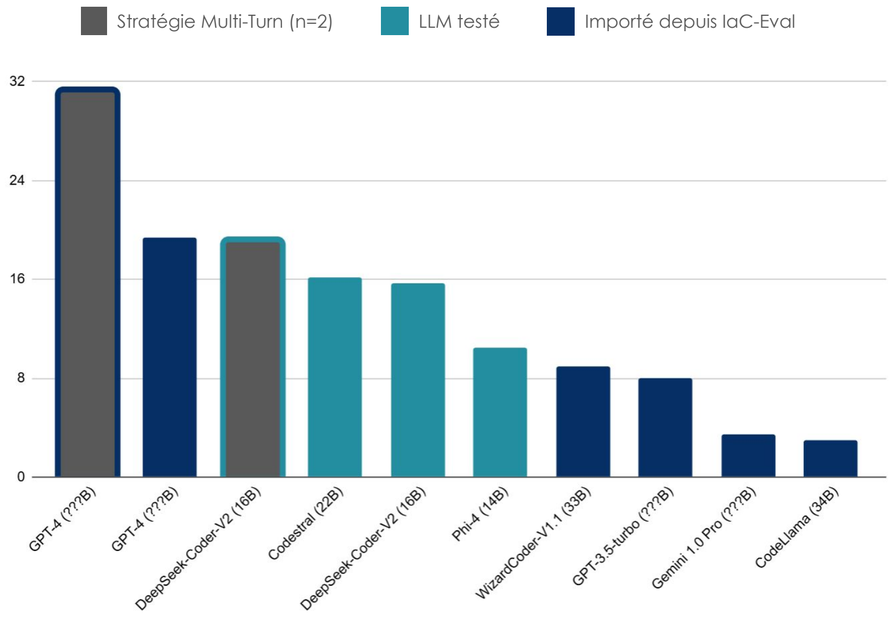
Workflow d'évaluation de IaC-Eval



Résultats

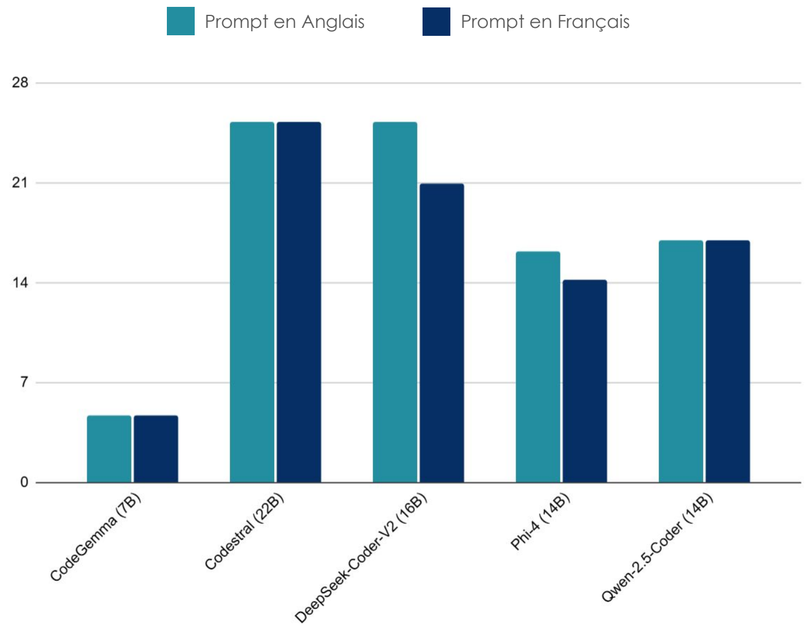
Performances des LLM sur l'ensemble du jeu de données :

Taux de réussite (%) - Niveaux 1, 2, 3, 4, 5, 6



Performances des prompts selon la langue utilisée (anglais vs français) :

Taux de réussite (%) - Niveaux 1, 2, 3 uniquement



En conclusion, **l'utilisation de prompts en français à la place de prompts en anglais a un impact relativement faible sur les taux de réussite**, voire nul selon les modèles. Avec DeepSeek-Coder V2 et Phi-4, la différence reste limitée, tandis qu'avec Codestral et Qwen-Coder 2.5, elle est inexistante.

Par ailleurs, la comparaison des performances des LLMs met en évidence que **certains modèles open source légers spécialisés en génération de code, tels que Codestral et DeepSeek-Coder V2, obtiennent des résultats très proches de GPT-4**, et surpassent nettement les autres modèles open source évalués dans le cadre d'IaC-Eval.

On remarque un facteur d'amélioration plus faible avec la stratégie Multi-turn pour DeepSeek Coder V2, en comparaison avec GPT-4.

Projet de recherche • Voie d'approfondissement Intégration et déploiement de systèmes d'information (VAP DSI) • 2024-2025
avril 2025

Contact pierre.laurent@telecom-sudparis.eu
adam.ouzegdouh@telecom-sudparis.eu

www.pierrelaurent.ovh/llm4cloud/