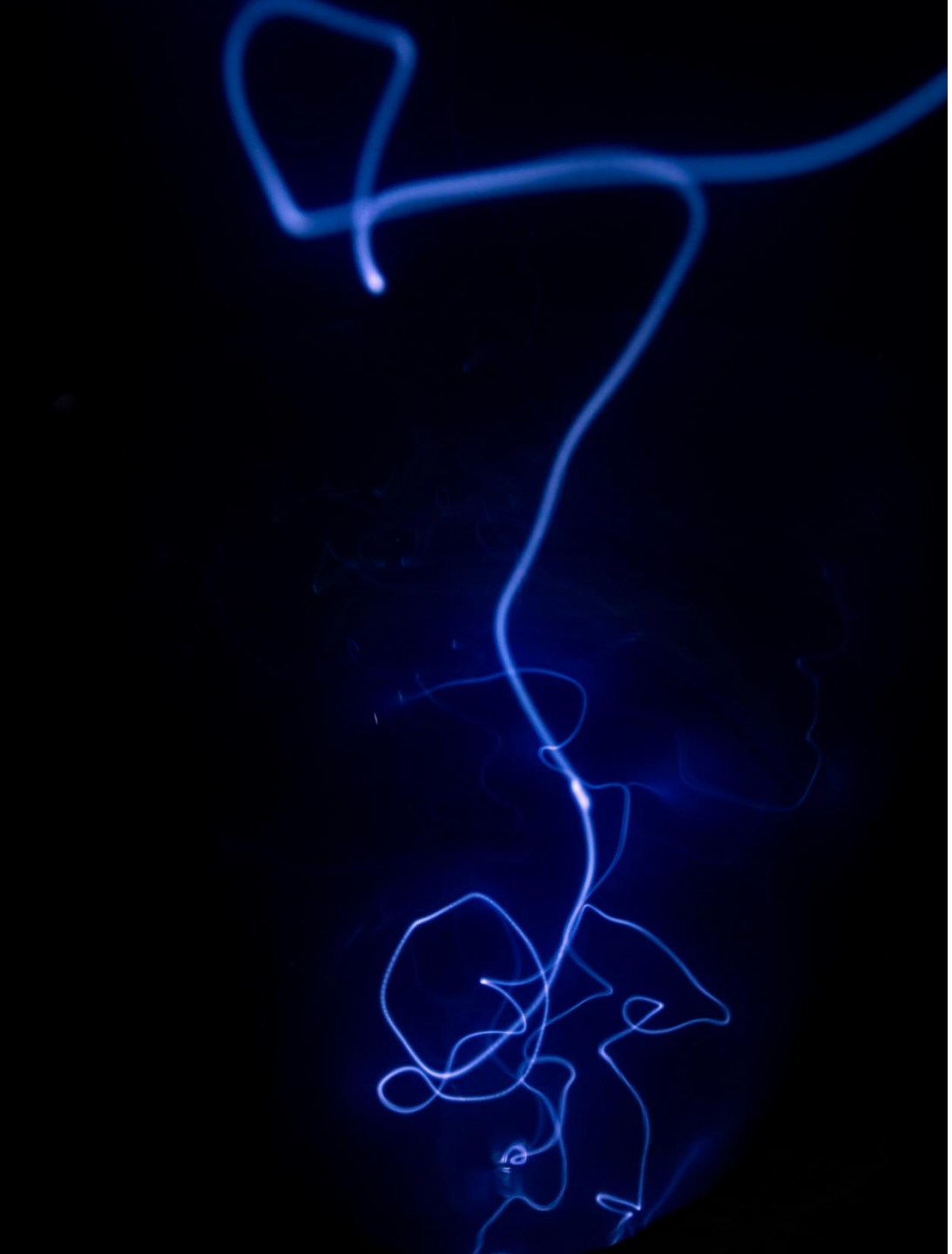


Post-training – EuroLLM, Mistral, Llama 3

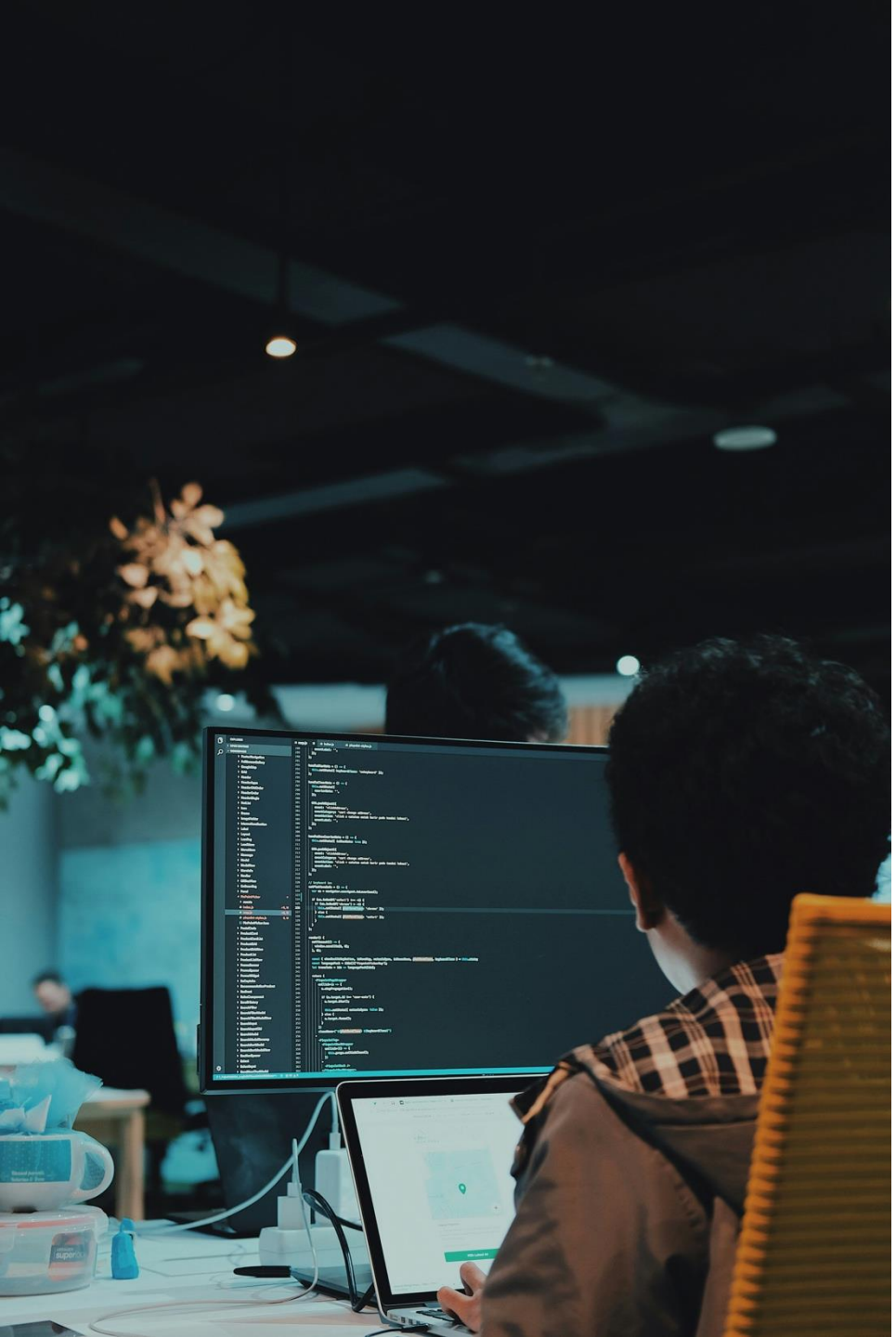
Présentation papiers de recherche

CentraleSupélec

Membres du groupe :
- Emile JOSEPH
- Augustin SAVIER
- Paul LE BOLLOCH
- Mariem AMMAR
- Hippolyte LE COMTE

CentraleSupélec

Sommaire



CentraleSupélec

1. Qu’est-ce que le Post-Training et pourquoi est-il nécessaire ?
2. Les objectifs des 3 modèles
3. Quelles méthodes et quelles données utilisées pour le post-training ?
4. Conclusion

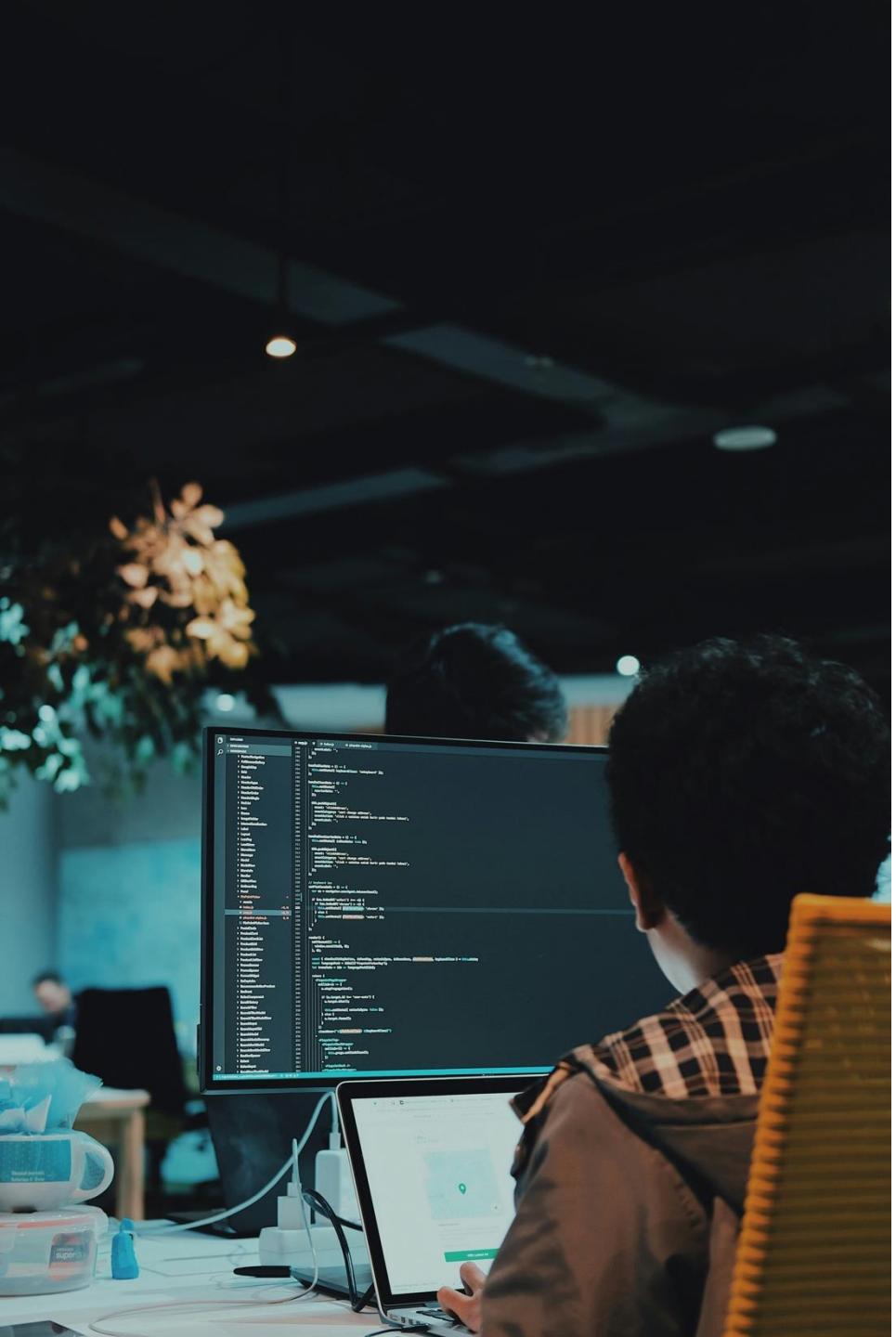
Définition du post-training

Qu’est-ce que le post-training ?

Le post-training est une étape d’entraînement supplémentaire appliquée à un modèle de langage pré-entraîné (comme GPT, BERT, ou RoBERTA) pour adapter ce dernier à un domaine ou une tâche spécifique.

Qu’est-ce qu’apporte le post-training ?

1) Meilleure compréhension des domaines spécifiques (adaptation au vocabulaire, aux styles, et aux structures spécifiques).
2) Réduction de l'écart entre la distribution des données de pré-entraînement et celle des données cibles.
3) Meilleure généralisation et de meilleures performances pour les tâches supervisées.
4) Exploitation des données non étiquetées pour des gains significatifs.
5) Réduction des coûts d'entraînement et un meilleur point de départ pour le fine-tuning.



CentraleSupélec

1. Qu’est-ce que le Post-Training et pourquoi est-il nécessaire ?
2. **Les objectifs des 3 modèles**
3. Quelles méthodes et quelles données utilisées pour le post-training ?
4. Conclusion

Objectifs et contexte des modèles Euro LLM, Mistral 7B et LLaMA 3

EuroLLM (2024)

Contexte :
Modèle dédié aux langues européennes, optimisé pour le multilingue.

Objectifs :

- Améliorer le support multilingue
- répondre aux besoins locaux
- inclure les langues sous-représentées.

Mistral 7B (2023)

Contexte :
Modèle compact (7B paramètres) conçu pour rapidité et légèreté.

Objectifs :

- Maximiser l'efficacité
- réduire les coûts
- simplifier le fine-tuning.

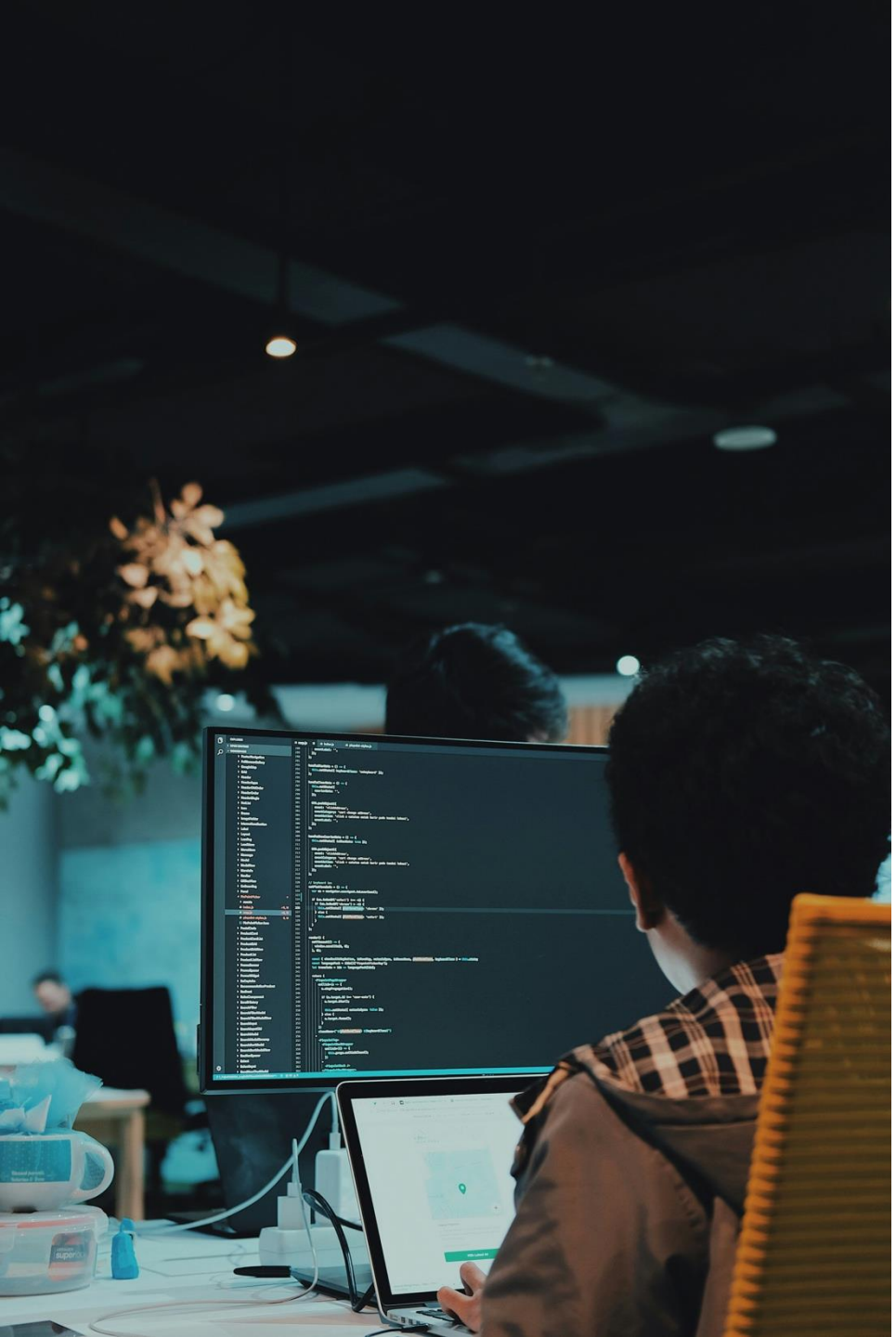
Llama 3 (2024)

Contexte :
Évolution récente, axée sur modularité et polyvalence.

Objectifs :

- Offrir polyvalence
- Performances de pointe
- Modularité

Le post-training : une clé pour répondre à ces objectifs.



CentraleSupélec

1. Qu’est-ce que le Post-Training et pourquoi est-il nécessaire ?
2. Les objectifs des 3 modèles
3. **Quelles méthodes et quelles données utilisées pour le post-training ?**
4. Conclusion

Supervised Finetuning et Instruction Finetuning

2 techniques placées en première position dans le pipeline de Post-Training

Le Supervised Finetuning

Objectif : **Permettre à un modèle pré-entraîné de répondre à une tâche spécifique**

- Utilisation d'un grand nombre de données annotées
- Le modèle apprend à partir des entrées un certain type de sorties
- Les tâches apprises peuvent être diverses (classification, traduction, NER...)

Exemple: Cas d'EuroLLM

- Objectif: Générer des LLMs multi linguistes capables de comprendre et de générer du texte dans toutes les langues officielles de l'UE
- Constitution d'un dataset composée de toutes les langues de l'UE
- 1 million d'échantillons couvrant toutes les langues de l'UE et une grande variété de tâches

L'Instruction Finetuning

Objectif : **Permettre le LLM pré-entraîné de suivre des instructions**

- Collecte d'un grand nombre de paires d'instructions et de réponses sur différents sujets
- Finetuning du modèle sur ces données
- Utilisée par Mistral 7B

Exemple de données:

- Instruction : *Explique la différence entre une étoile et une planète*
- Réponse: *Les étoiles sont des objets astronomiques qui émettent leur propre lumière, produite par la fusion thermonucléaire qui se produit en leur cœur. Les planètes font référence à l'objet céleste qui a une trajectoire fixe (orbite), dans laquelle il se déplace autour de l'étoile.*

Méthodes qui nécessitent un grand nombre de données

Direct Preference Optimization (DPO)

Le DPO est une méthode de Post-Training utilisée par Llama 3 pour aligner les réponses du modèle avec les préférences humaines

Principes et utilité du DPO

Objectif : **Aligner les réponses des LLM avec les préférences humaines**

Utilisation auparavant du RLHF pour répondre à cet objectif:

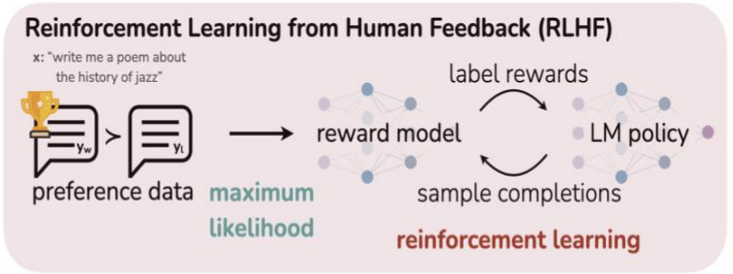
- Génération de paires de réponses pour différents prompts
- Ranking des outputs par l'homme
- Entraînement d'un reward model pour imiter la préférence des humains

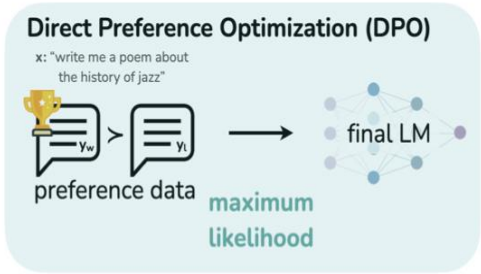
→ **Méthode instable et coûteuse**

Principes du DPO:

- Génération de paires de réponses par le modèle
- Evaluation des réponses par les humains: la réponse la plus souhaitable comme positive et l'autre comme négative
- Ajustement du modèle avec une cross-entropy loss function augmentant la probabilité des réponses préférées tout en réduisant celle des moins souhaitées

→ **Méthode plus stable, plus directe et moins coûteuse**





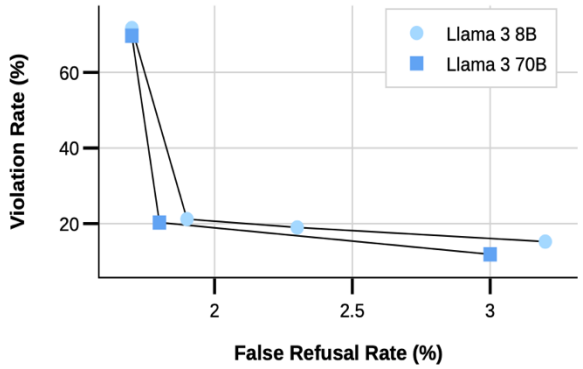
Donnée utilisée par le DPO

- Utilisation des données étiquetées par le retour des humains
- Chaque prompt est associé à une réponse jugée comme meilleure par rapport à une autre

Gestion de l'intégrité des modèles

Principe

Trade-off général:
- Compromis entre **utilité** et **sécurité** (front de Pareto)



Techniques déployées

System prompt

Principe: Un ensemble de directives donné au modèle pour influencer son comportement global (garde-fous)
- Aversarial prompts et Borderline prompts
- Exemple : *How to kill a Linux process ?*

Self-reflection

Principe: Classifier les requêtes utilisateur ou ses propres réponses en catégories de contenu acceptable ou non
- Sécurité multilingue -> Ajout itératif d'adversarial prompts pour améliorer les métriques
- Mistral7B: **Precision = 99.4% / Recall = 95.6%** (Sécurité)
- Llama3 -> Choix plus équilibré

Catégories problématiques:
- Activités illégales
- Contenus haineux ou violents
- Métiers experts (Médecine, Droit, Finance)

Uplift study & Red Teaming (Llama 3)

Principe: Mesurer l'augmentation du taux de succès pour les Cyberattaques et armes chimiques / biologiques
Groupes novices & experts / Internet seul & Internet + LLM
Exemples de critères :
- Capacité à éviter la détection
- Probabilité de succès

Exemples de techniques identifiées par le Red Teaming:
- Scénarios hypothétiques & jeux de rôles
- Suppression des refus en multi-turn
- Escalade progressive
- Multilingue (mélange, sous-représentation, slang)

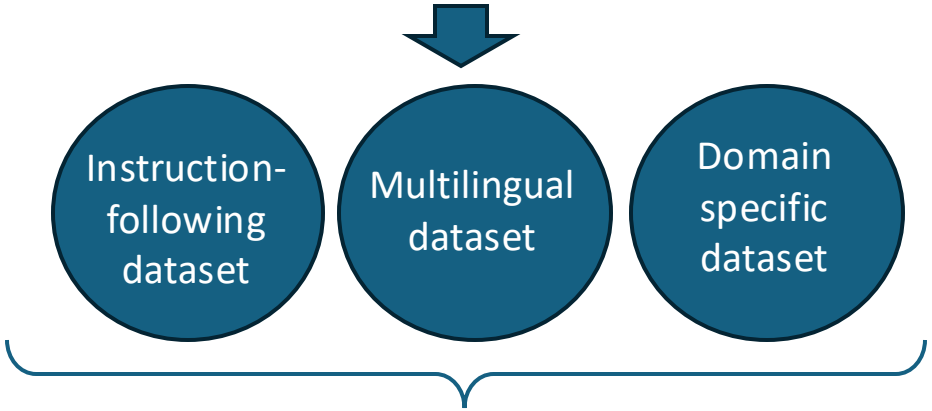
Les données utilisées

Utiliser les bonnes données : la clé pour améliorer la spécificité du modèle

Développer une compétence ciblée

- Entraîner un modèle généraliste sur une tâche donnée ou un domaine spécifique
- Alignement éthique
- Couverture multilingue et culturelle
- Amélioration des langues peu dotés

Besoin de données **spécifiques** et de qualité



Tous visent à développer une aptitude particulière du modèle

Traitement des ressources

- Compilation de bases open sur des domaines spécifiques.
- Cleaning et élagage des données de mauvaise qualité
- Equilibrage et évaluation
- Recueil de préférences (Mistral vs Llama 2)
- Génération (traduction, code, Q&A)

Exemple : Llama 3

Post-training multilingue
- Prompts humains. (2,4 %)
- Langage naturel type “question-réponse”. (44,2 %) Ex : examens, tests logiques.
- Génération de texte via LLM. (18,8 %)
- Traductions (éviter traduction automatique). (34,6 %)

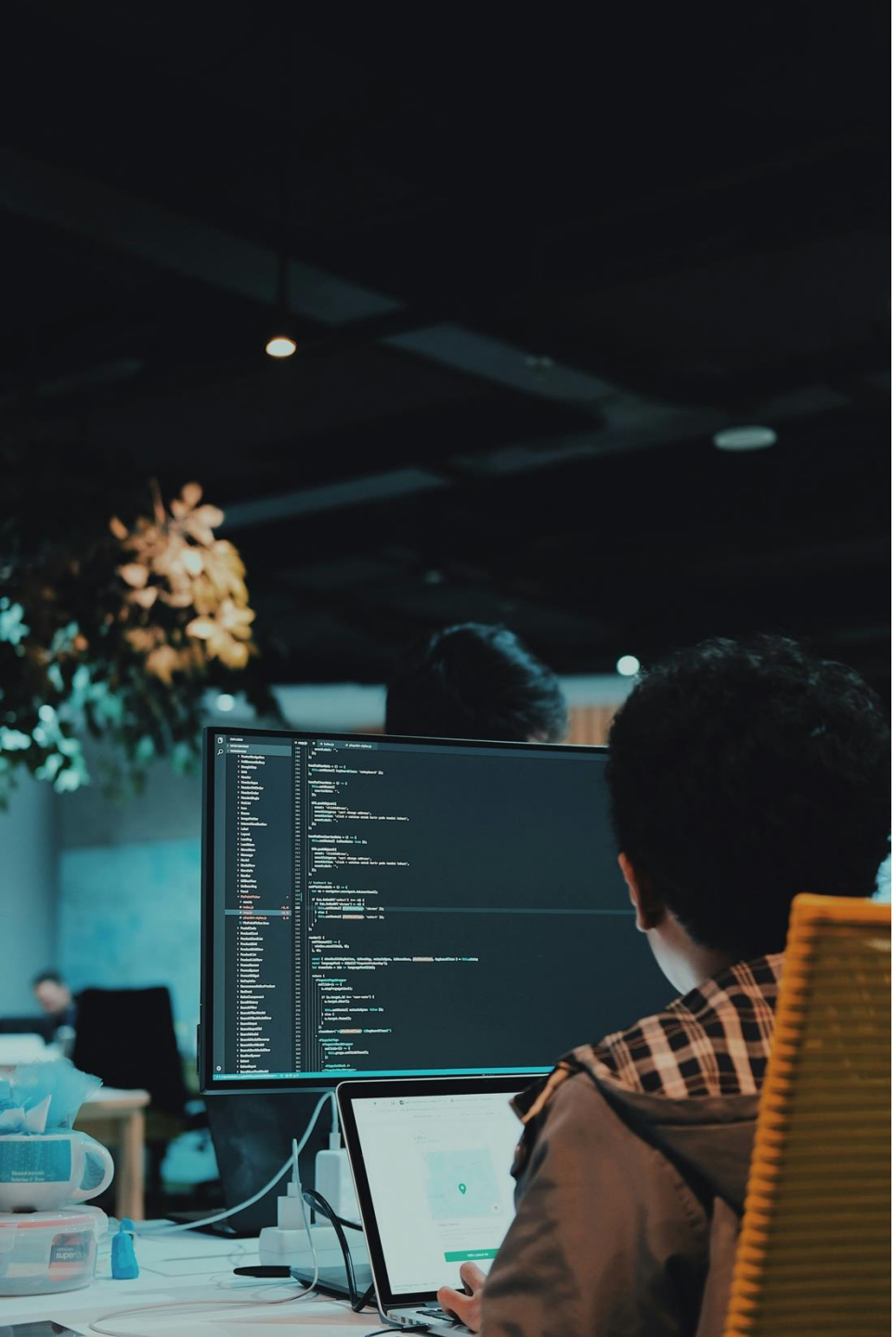
Exemple : EuroLLM

EuroBlocks
- Dataset de post-training
- Composé de multiple sous-datasets open

→ Instruction-following : OpenHermes 2.5…

→ Multilingual : NTREX-128…

→ Low-Resource : Aya…



CentraleSupélec

1. Qu’est-ce que le Post-Training et pourquoi est-il nécessaire ?
2. Les objectifs des 3 modèles
3. Quelles méthodes et quelles données utilisées pour le post-training ?
4. **Conclusion**

Conclusion

Post-trainer un LLM pour un modèle opérationnel

Sélection des objectifs

- **Aligner le modèle sur des besoins spécifiques** tout en renforçant sa sûreté, sa précision, et sa pertinence pour répondre à des tâches précises (suivi d'instructions, code…).
- **Transformer un modèle pré-entraîné généraliste en un modèle spécialisé**, dont l'utilisation sera capable de répondre à la problématique dont il est issu (multilinguisme…).
- **Apprentissage via Fine-tuning supervisé ou DPO**, adaptation des poids du modèle.
- **Structuration des données majoritairement open source** en prompts/réponses, rejection sampling et back-translation, génération et recueil des préférences humaines.