

StarCoder 2 and The Stack v2: The Next Generation

Built by BigCode

Équipe CentraleSupélec

Nancy Abi Farah	nancy.abifarah@student-cs.fr
Irene Dagher	irene.dagher@student-cs.fr
Noémie Guisnel	noemie.guisnel@student-cs.fr
Anthony Quentin	anthony.quentin@student-cs.fr
Martin Rampon	martin.rampon@student-cs.fr



StarCoder 2 and The Stack v2: The Next Generation
The BigCode project, an open-scientific collaboration focused on the responsible development of Large Language Models for Code (Code LLMs), introduces StarCoder2. In partnership with Software...
[arXiv.org](https://arxiv.org)

Sommaire

1 - Introduction

2 - Évaluation du modèle

3 - Dataset : The Stack

4 - Pipeline du prétraitement des données

5 - Modèle : StarCoder

6 - Impact social, limitations et challenges

1 - Introduction

Qu'est ce que BigCode?



- collaboration scientifique ouverte
- créée en 2022
- soutenue par servicenow et hugging face
- **utilisation responsable** de LLM

transparence

en partageant ouvertement les ensembles de données, les modèles et les expériences

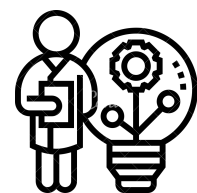
gouvernance

un contrôle sur la gestion de leur propriété intellectuelle

collaboration

chercheurs en IA et membres de la communauté open source

 **bigcode/****in-the-stack**  **Am I in the Stack?**



OptOut

Stay in the stack

Code retiré ou pas de l'ensemble de donnée

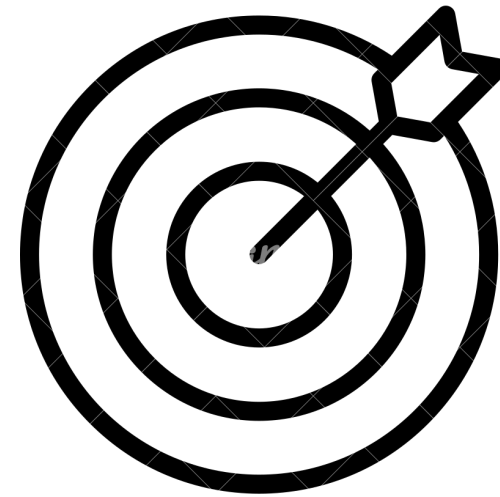
1 - Introduction

Les enjeux du papier



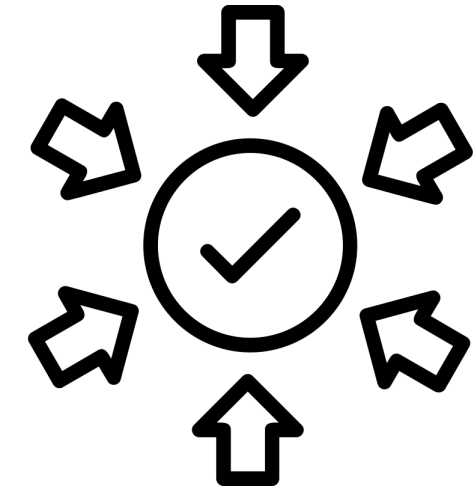
Problème

- Manque de transparence autour de la création des LLMs
- Manque de transparence sur les données (D'où est ce qu'elles proviennent? Comment sont-elles prétraitées?)



Objectifs

- Entraîner un modèle, qui a de bons résultats et partager en détails la provenance données et les traitements faits dessus



Implications

- Renforcer la confiance dans les modèles développés

1 - Introduction

Explication *rapide* du papier



Structure du papier

Introduction et enjeux
Source de données
Pipeline du prétraitement
Formatage des données
Architecture du modèle et détails d'entraînement
Évaluation du modèle
Impact sociaux, limitations, challenges et risques

The STACK v2 

StarCoder2 
transformer

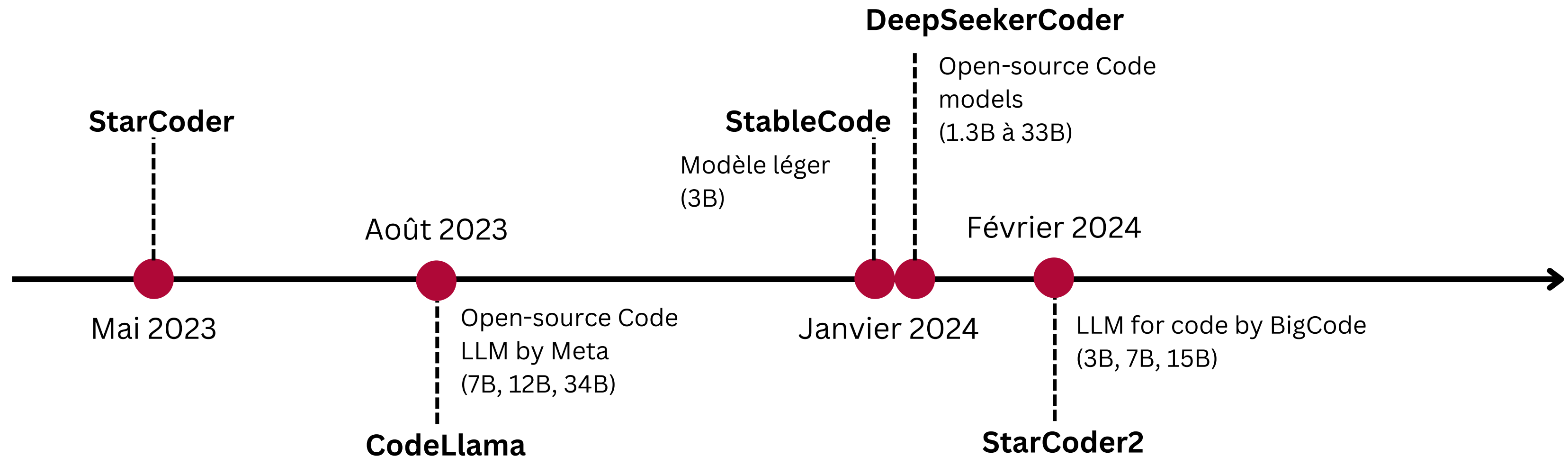
Objectifs du papier

- Présenter en détails STACK2  pour entrainer les LLMS
- Présenter le modèle StarCoder2 : modèle de pointe
- Renforcer la confiance dans le modèle développé
- Concilier IA performante et éthique

1 - Introduction

Timeline du papier :

état de l'art actuel des modèles open-source pour le code



2 - Évaluation du modèle

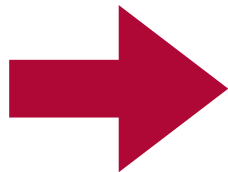
Comparaison à nombre de paramètres équivalent avec les modèles déjà existants **StableCode**, **Code Llama**, **DeepSeekCoder**

Tâches évaluées :

- Complétion de code
- DataScience avec Python
- Correction et édition du code
- Raisonnement mathématiques

Table 9: Pass@1 on HumanEval(+) and MBPP(+). These results were generated using greedy decoding.

Model	HumanEval	HumanEval+	MBPP	MBPP+
StarCoderBase-3B	21.3	17.1	42.6	35.8
DeepSeekCoder-1.3B	28.7	23.8	55.4	46.9
StableCode-3B	28.7	24.4	53.1	43.1
StarCoder2-3B	31.7	27.4	57.4	47.4
StarCoderBase-7B	30.5	25.0	47.4	39.6
CodeLlama-7B	33.5	25.6	52.1	41.6
DeepSeekCoder-6.7B	47.6	39.6	70.2	56.6
StarCoder2-7B	35.4	29.9	54.4	45.6
StarCoderBase-15B	29.3	25.6	50.6	43.6
CodeLlama-13B	37.8	32.3	62.4	52.4
StarCoder2-15B	46.3	37.8	66.2	53.1
CodeLlama-34B	48.2	44.3	65.4	52.4
DeepSeekCoder-33B	54.3	46.3	73.2	59.1

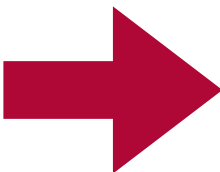


StarCoder2-15B surpasse les autres “grands” modèles dans la grande majorité des tâches

2 - Évaluation du modèle

Comparaison à nombre de paramètres équivalent avec les modèles déjà existants **StableCode**, **Code Llama**, **DeepSeekCoder**

StarCoder2	Tâches						
	Code Completion			Code fixind and editing		Math Reasoning	Fill-in-the-Middle
	HumanEvall MBPP	MultiPL-E (18 langages)	Data Science	HumanEvalFix (6 langages)	CanItEdit	GSM8K (middle school)	FIM
3B	1 ^{er}	1 ^{er} sur 11/18 des langages	1 ^{er}	×	2 ^e	2 ^e derrière la v1	×
7B	2 ^e	DeepSeekCoder	2 ^e	×	2 ^e	2e derrière DSC	×
15B	1 ^{er}	1 ^{er} sur 16/18 des langages	1 ^{er}	2 ^e derrière DSC	1 ^{er}	1 ^{er}	3 ^e



StarCoder2-15B surpasse les autres “grands” modèles dans la grande majorité des tâches

3 - Dataset The Stack V2

The Stack V2

Ensemble de données de pré-entraînement pour les “Code LLMs”

- Amélioration de The Stack V1
- Plus de 3 milliards de fichiers
- Plus de 600 langages de programmation
- Opensource

	The Stack v1	The Stack v2
full	6.4TB	67.5TB
dedup	2.9TB	32.1TB
train (full)	~200B tokens	~900B tokens

Comparaison de la taille de The Stack V1 et V2
<https://huggingface.co/datasets/bigcode/the-stack-v2>

➡ **"Am I in The Stack"**: permet aux développeurs de vérifier si leur codes (par exemple sur github) sont dans le dataset

3 - Dataset The Stack V2

The Stack V2

En collaboration avec
Software Heritage



- Plus grande archive publique mondiale de **codes sources**
- Composé de **104.2M** de dépôts github
- Application de **filtres** pour supprimer les fichiers de **mauvaise qualité** (fichiers de données, fichiers générés automatiquement)



Github issues
Pull Requests



Documentation sur
des gestionnaires de
packages



Notebooks



Regroupement de
petits datasets de
maths et de code



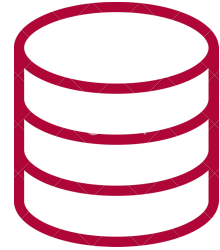
WIKIPEDIA

arXiv

Autres datasets
de NLP

4 - Pipeline de Prétraitement des données

Structure et étapes Clés



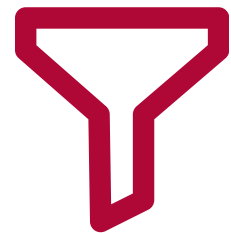
Collecte des données

Sources multiples : Software Heritage, GitHub, Kaggle, StackOverflow, etc.



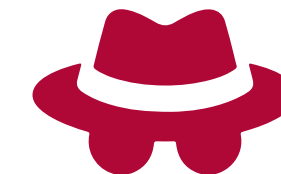
Déduplication des données

Déduplication et élimination des doublons avec des techniques avancées (LSH).



Filtrage des données

Suppression des fichiers autogénérés, des données mal formées, et des fichiers contenant peu de contenu pertinent.



Anonymisation et confidentialité

Utilisation de modèles comme StarPII pour détecter et anonymiser les noms, emails, clés, adresses IP, mots de passe, et noms d'utilisateur.

4 - Pipeline de Prétraitement des données

Étapes avancées et outils spécifiques

- **Détection des logiciels malveillants :** 59 442 fichiers malveillants supprimés (0,009 % des fichiers).

ClamAV est utilisé pour scanner et supprimer les malwares dans le code.

- **Suppression des benchmarks contaminés :** Élimination des données qui pourraient chevaucher les données de test.

Nettoyage des fichiers pouvant biaiser les évaluations (HumanEval, GSM8K).

- **Gestion des demandes de retrait (Opt-out):**

Utilisation de l'outil "Am I in The Stack" pour permettre aux développeurs de vérifier si leur code a été inclus dans le dataset.

- **Filtrage contextuel :**

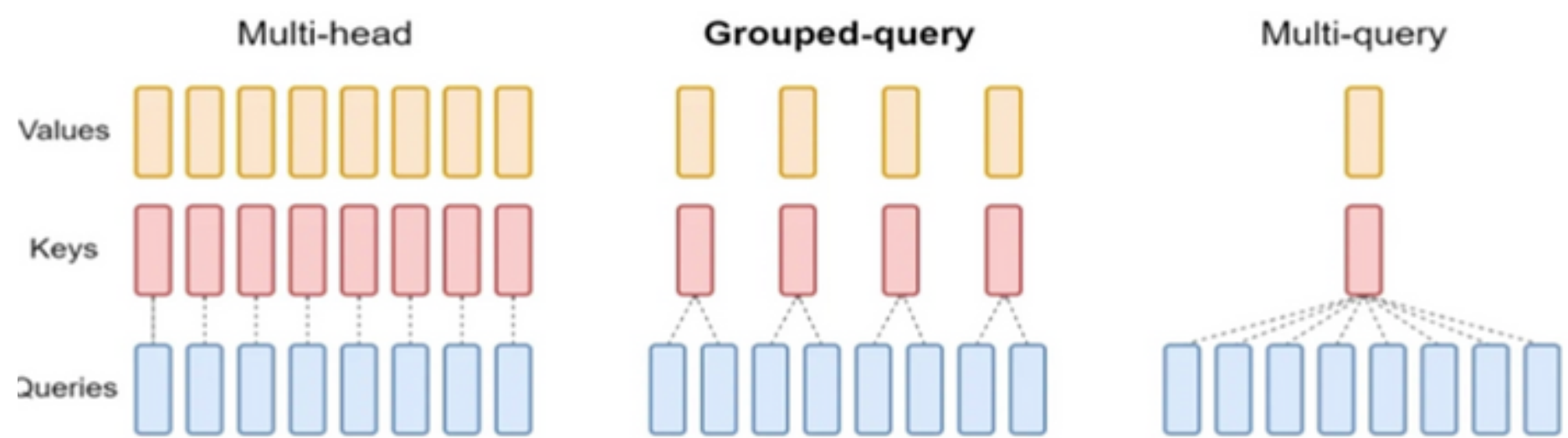
i. **Filtres de base :** Suppression des fichiers avec plus de 100 000 lignes, des lignes >100 caractères, ou générés automatiquement.

ii. **Filtres spécifiques à la langue :** Actions ciblées pour des formats comme HTML, JSON ou Markdown pour améliorer la qualité.

5 - Modèle : StarCoder2

Spécificités de StarCoder2

Grouped-query attention



Rotary Positional Encoder

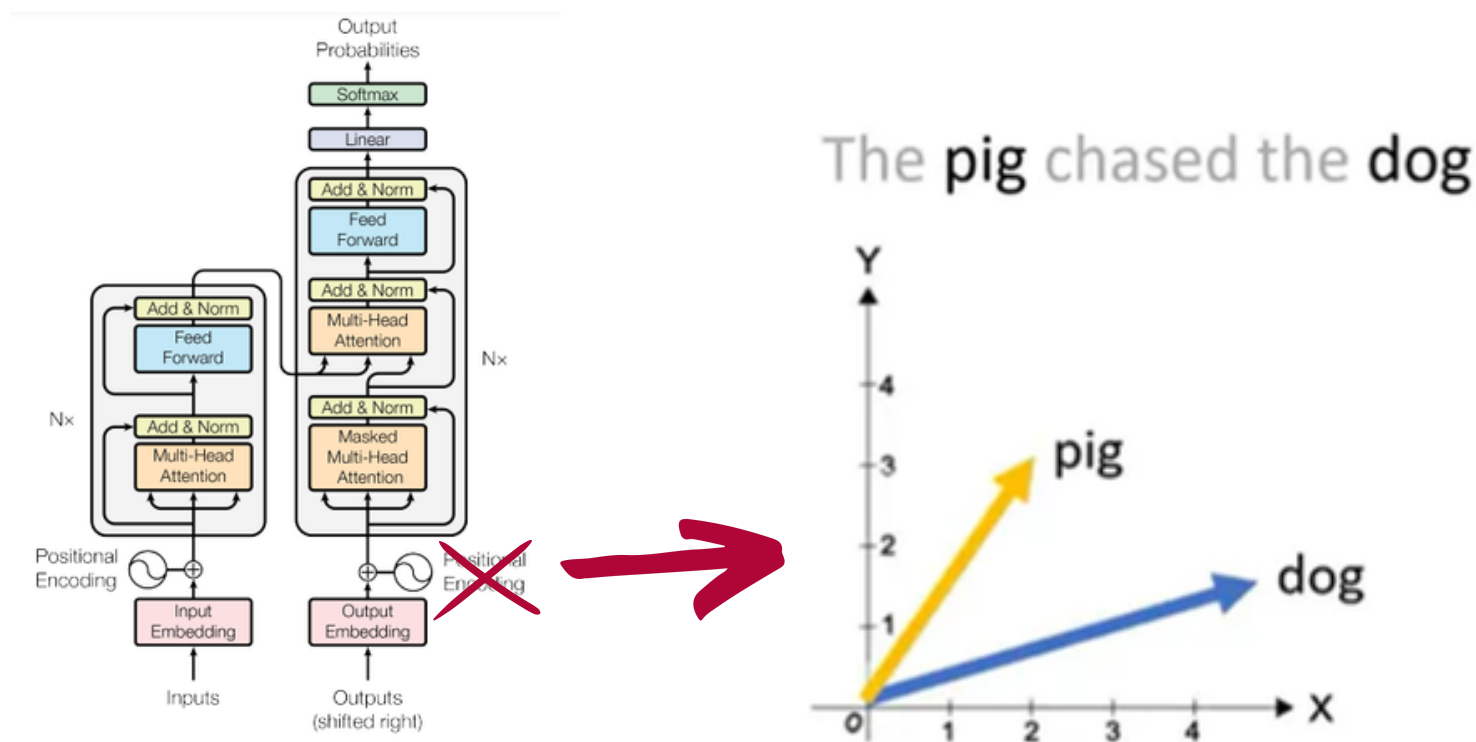


Figure 1: The Transformer - model architecture.

	Multi-Head	Grouped Query	Multi-Query
Avantages	<ul style="list-style-type: none">- Spécialisation des têtes.- Représentation riche / diversifiée.	<ul style="list-style-type: none">- Équilibré- Optimisation mémoire + calculs groupés	<ul style="list-style-type: none">- Inférences rapides : une seule clé/valeur.- Haute efficacité GPU.
Inconvénients	<ul style="list-style-type: none">- Consomme beaucoup de mémoire et de calcul.- Lent pour les inférences rapides.	<ul style="list-style-type: none">- Complexité (configuration des groupes)- Dépend des groupements.	<ul style="list-style-type: none">- Perte de diversité dans les réponses.- Moins performant pour les représentations complexes.
Exemple d'utilisation	GPT-3 (OpenAI)	StarCoder v2	LLaMA (Meta)

- **Principe:** code les relations via matrice de rotation
- **Avantage :** Améliore généralisation des modèles (séquences longues)
- **Application :** Meilleure gestion de la dépendance entre positions dans des contextes variés

6 - Éthique

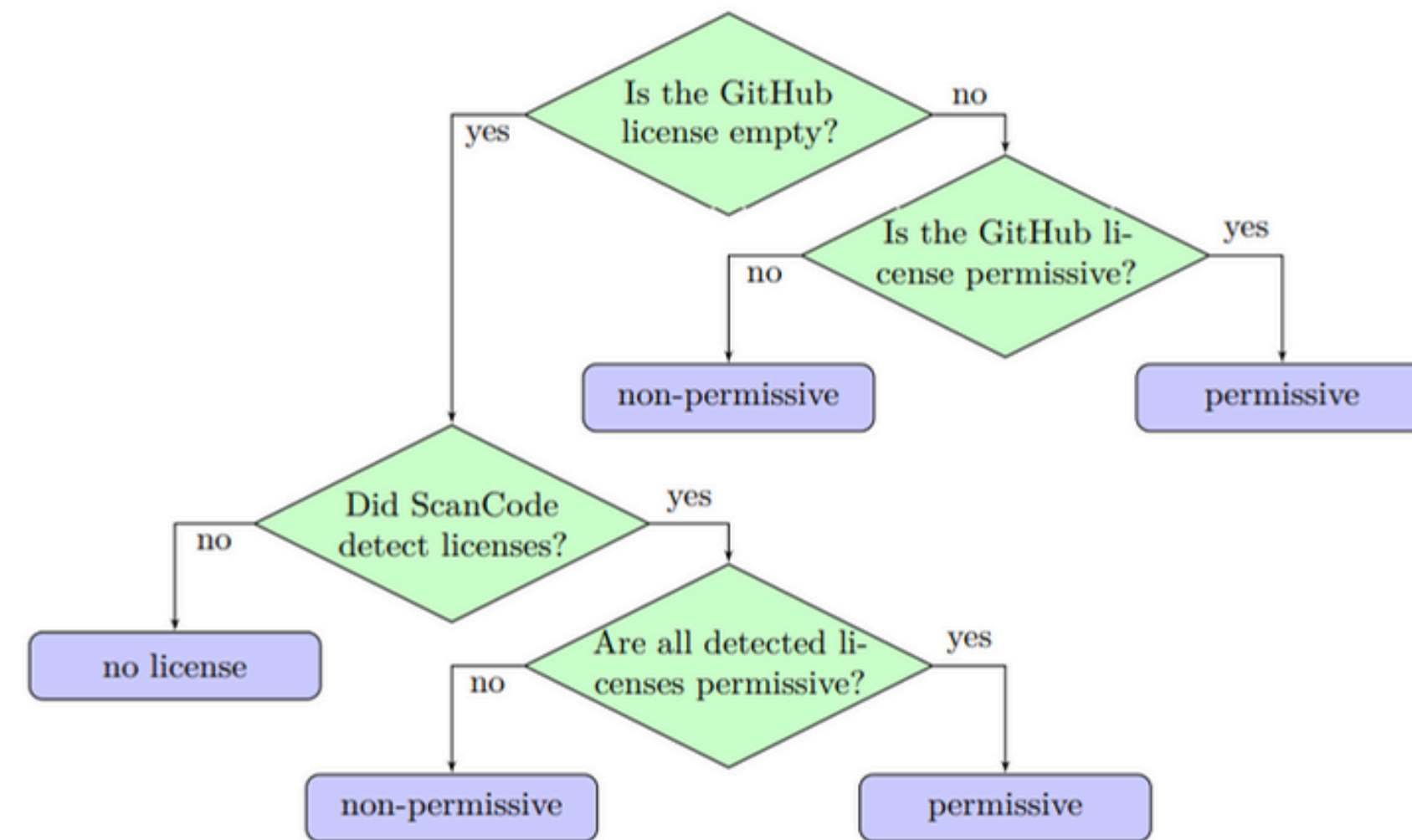
Personally Identifiable Information (PII)



- **Rédaction PII:** Suppression des noms, emails, mots de passe, etc.
- **Sources ciblées:** Code, PRs, StackOverflow, arXiv.
- **Anonymisation:** Pseudonymes remplacés par des compteurs (username_1, username_2) -> conserve les échanges

Outil : StarPII

Licenses



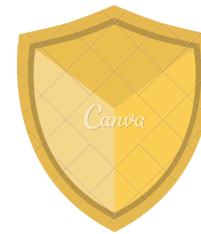
96.93% des repos GitHub : pas de license

6 - Limitations / Challenges



Confidentialité

Gestion complexe des PII dans le code généré



Sécurité

Artefacts exploitables par des acteurs malveillants



Biais sociaux

Peut renforcer des biais sociaux (stéréotypes, genres...)



Emplois

L'automatisation du code supprime et crée des opportunités

Traçabilité

Difficulté à tracer les composants logiciels (ex. SWHID)

Biais de représentation

Favorise l'anglais et langages populaires

