

AUTHOR

Bogdan-Mihai Buzatu
b.m.buzatu-1@student.tudelft.nl

AFFILIATIONS

Delft University of Technology
Delft, the Netherlands

DATA HOUND: ANALYSING NON-ENGLISH DATA SMELLS IN LARGE CODE DATASETS

SUPERVISORS

Prof. Dr. Arie van Deursen
Assistant Prof. Dr. Maliheh Izadi
Ir. Jonathan Katzy
Ir. Razvan-Mihai Popescu



INTRODUCTION

Large-language models (LLMs) has been increasingly used for coding purposes. Often they are trained on datasets that contain **“Uneven Natural Languages” data smell** [1][2], non-English snippets embedded data. We build a three-stage pipeline (Detection → Generation → Evaluation) that tags every character with its predicted language, masks selected spans and probes three open-source models (SmolLM2 135 M, StarCoder 2 15 B, Mellum-4 B) on the Java portion (3.35 M files) of **The Heap** [3]. Our aim is to determine how the presence or removal of non-English tokens affects **code generation and summarisation** quality.

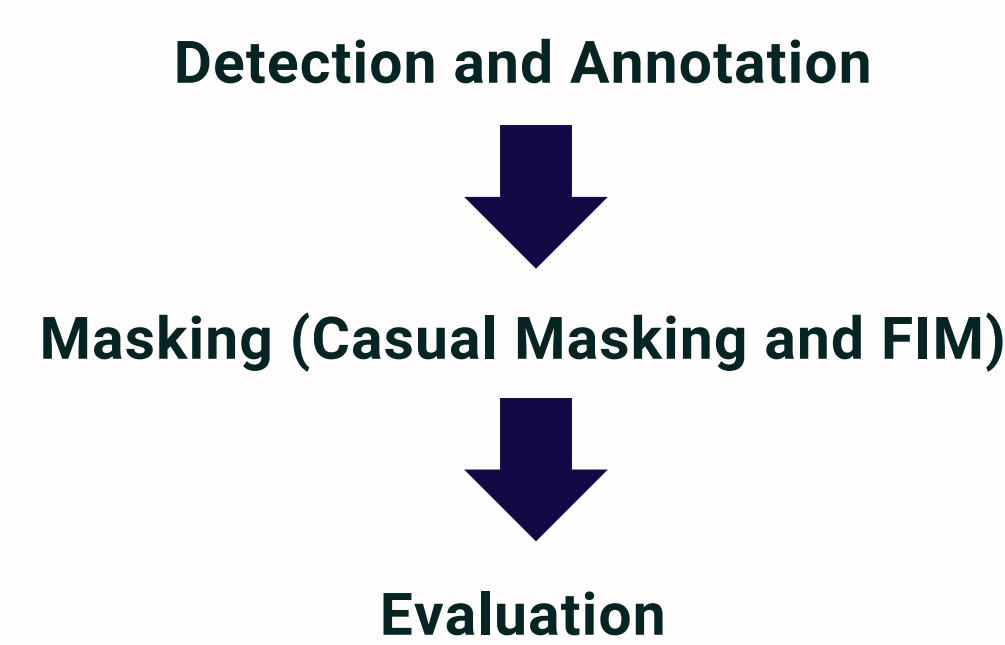
RESEARCH QUESTIONS

- RQ1** What languages are commonly used in the code in addition to English?
- RQ2** What is the distribution of English and non-English elements across the Heap?
- RQ3** What is the effect on Large Language Models’ code generation when non-English code elements are present in the prompt dataset?
- RQ4** How is Large Language Models’ code summarisation influenced when non-English code elements are masked in the prompt dataset?

RELATED LITERATURE

- [1] Harald Foidl, Michael Felderer, and Rudolf Ramler. Data smells: Categories, causes and consequences, and detection of suspicious data in ai-based systems, 2022.
- [2] Antonio Vitale, Rocco Oliveto, and Simone Scalabrino. A catalog of data smells for coding tasks. ACM Trans. Softw. Eng. Methodol., December 2024. Just Accepted.
- [3] Jonathan Katzy, Razvan Mihai Popescu, Arie van Deursen, and Maliheh Izadi. The heap: A contamination-free multilingual code dataset for evaluating large language models, 2025.
- [4] Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle, 2022.

METHODOLOGY

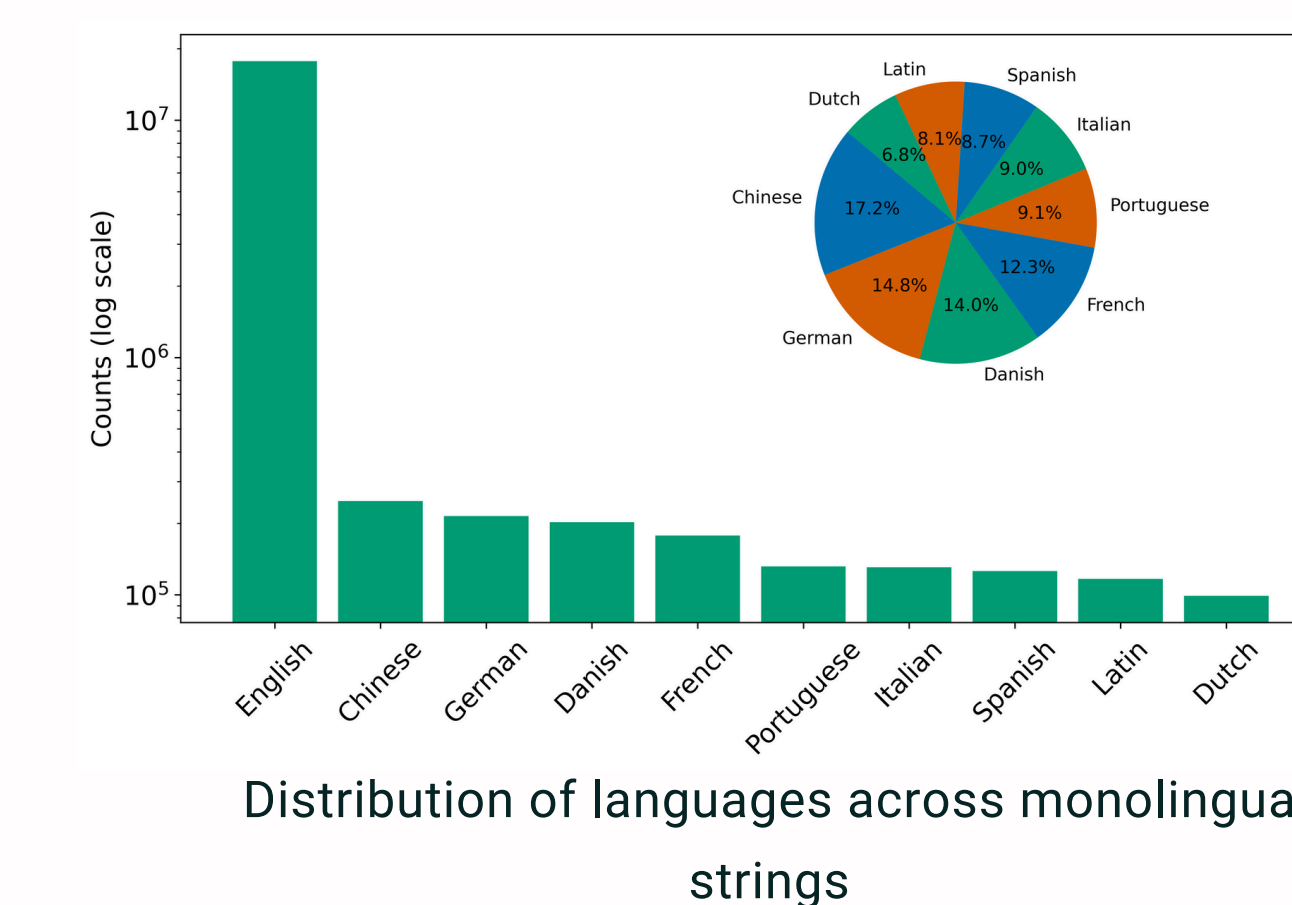
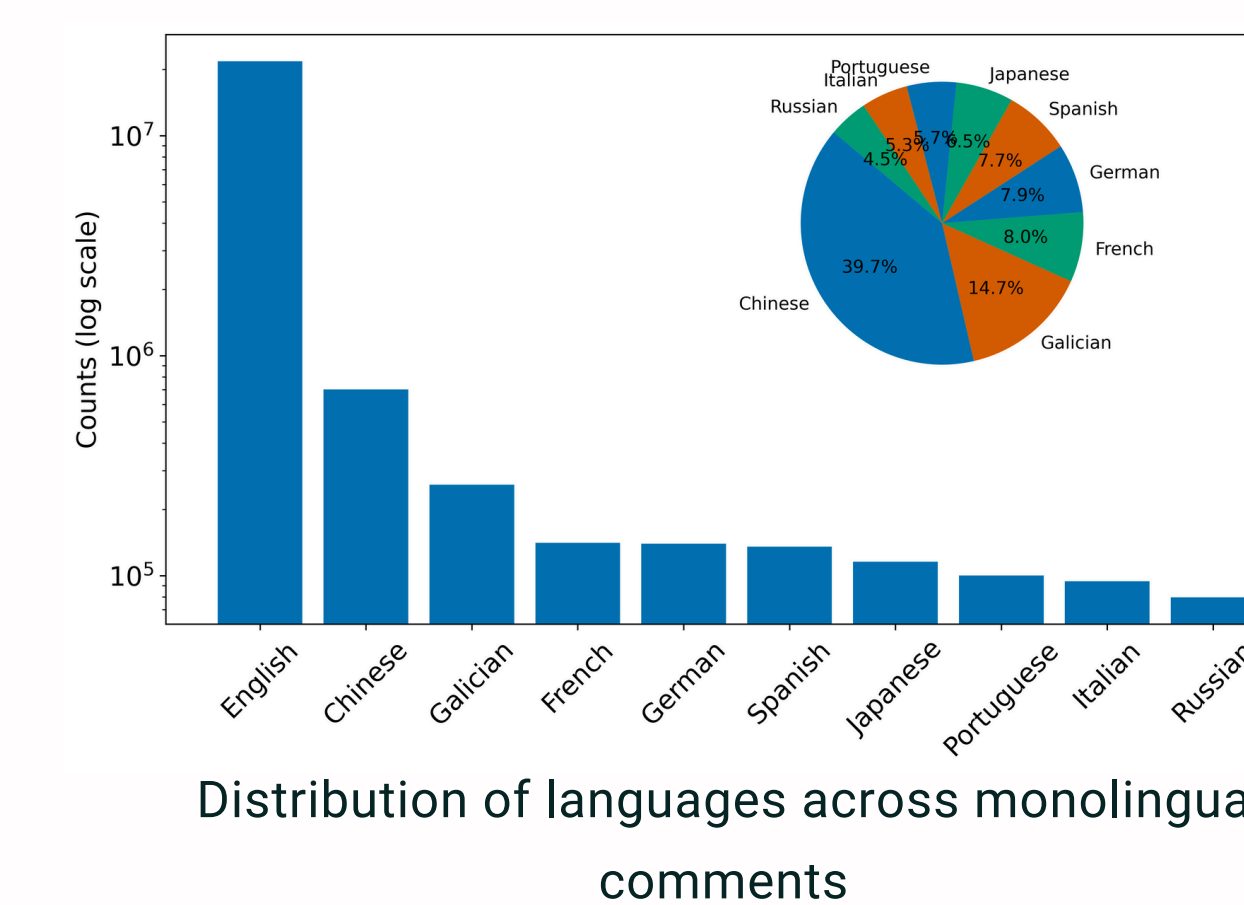
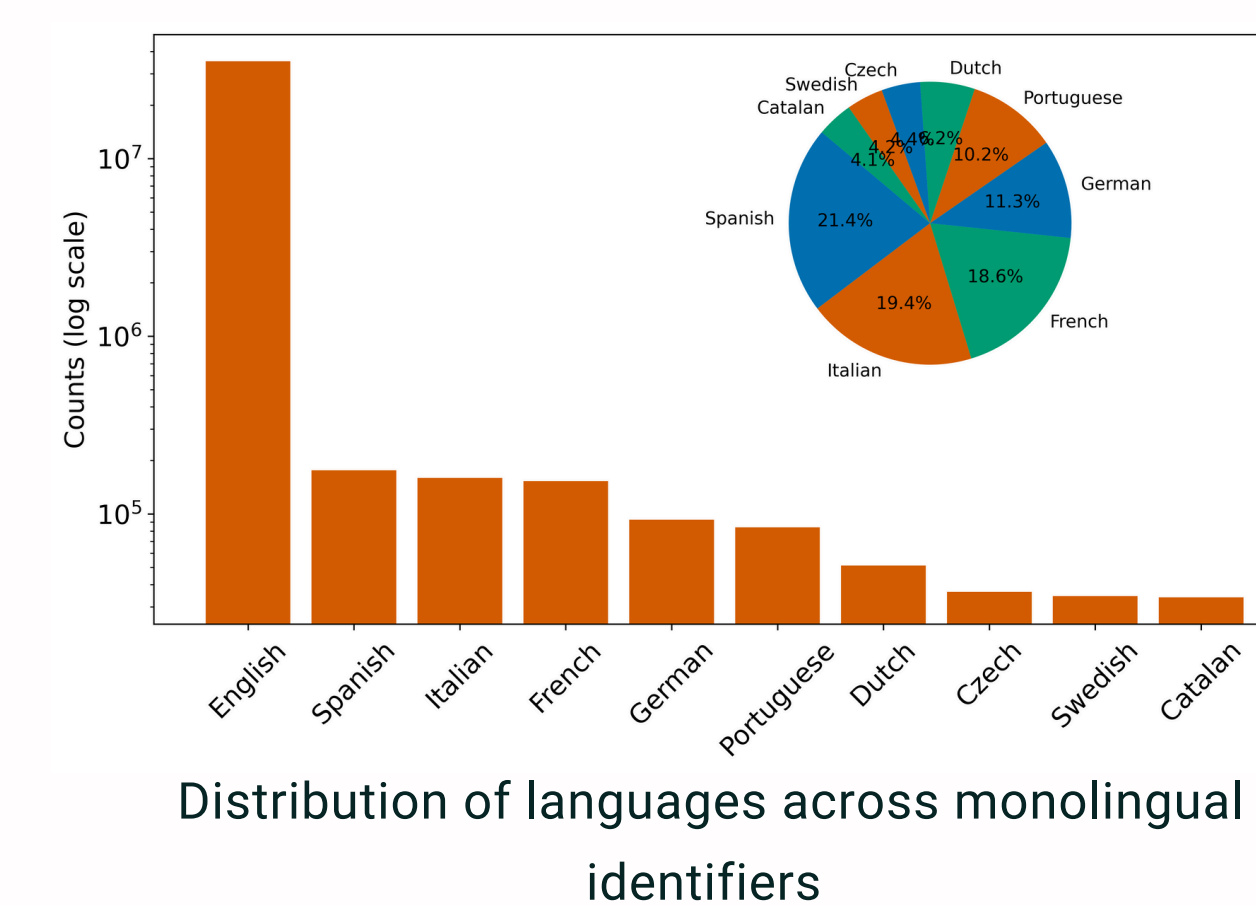


For the detection, we use pyculd2 and FastText. After character-level annotation, we split the resulting dataset into English-only and non-English subsets, sample 1 000–2 000 files each, and apply causal masking (next-line prediction) and Fill-in-the-Middle masking [4] (docstring/block-comment infill). The masked contexts are fed to SmolLM2, StarCoder 2 and Mellum. Outputs are scored with Exact Match, BLEU-4, METEOR, and ROUGE-1/-2/-L to capture surface overlap, fluency and recall.

RESULTS/FINDINGS

- English still dominates**, accounting for > 90 % of tokens across comments, strings and identifiers; Chinese, Spanish, Portuguese and French form a long-tailed minority.
- Non-English helps**: BLEU, METEOR and ROUGE scores rise slightly whenever non-English elements are present or masked, suggesting that shorter, more formulaic patterns in those languages are easier for models to reconstruct.
- Model hierarchy is stable**: Mellum yields the most fluent continuations, StarCoder 2 recalls the broadest vocabulary, and SmolLM2 trails on both axes—regardless of language mix.
- Masking matters**: Hiding non-English spans further boosts Mellum’s BLEU while StarCoder 2 preserves recall (ROUGE-1/L); both models regenerate non-English comments more faithfully than English ones.

ANALYSIS



CONCLUSION

- We provide a **reproducible Detection→Generation→Evaluation** pipeline with public code and character-level language annotations.
- Non-English tokens, **though rare, systematically seems to perform better in the evaluation metrics**, revealing subtle biases and opportunities for multilingual tuning.

