6 rue du coq Français
Les Lilas, 93260
(+33) 6 59 90 85 14
nathanael.beau.gs@gmail.com
NathanaelBeau

Nathanaël BEAU

Curriculum Vitæ

Educational background

2020–2024 **Ph.D. in NLP : Generating Python code from a Natural Language description**, *Université de Paris Cité*, Paris

Subject : The central hypothesis of this thesis investigates the naturalness hypothesis, which posits that software functions as a medium of human communication, shares statistical characteristics similar to those of natural language corpora, and that these characteristics can be leveraged to develop superior software engineering tools. The objectives of this research are to explore this hypothesis and investigate how NLP techniques can be applied to enhance the functionality and usability of code generation systems.

2016–2019 **Arts et Métiers ParisTech - Master's degree : Master of Science In Engineering**, Paris, National ranking : 356/3913

Master in Machine Learning for the third year. Relevant courses : mathematics, machine learning, operations research, project management, value engineering, supply chain, electronics, mechanics, industrial engineering.

2013–2016 **Undergraduate Program / Classes préparatoires aux Grandes Ecoles**, *Lycée Charlemagne*, Paris

Relevant courses : mathematics, physics, chemistry, engineering sciences, literature and foreign languages.

Work experience

2020– 2023 **PhD Student - CIFRE Program**, *onepoint*, Paris, France

Conducted doctoral research in collaboration with onepoint under the CIFRE program, bridging academic research and industrial application.

Activities :

- Implemented and trained the RETROcode model, a RAG model for development assistance, achieving competitive results with top LLMs such as Codex and Mistral.
- Developed and implemented the BertranX model to ensure syntactic validity of generated code.
- Led a team of 5 data science experts to create the CodeInsight dataset for development aid.
- Delivered customer training on large language models, including a recurrent training session for Chanel on ChatGPT.
- Supervised a work-study student for two years, overseeing projects on comparing Seq2seq and Language Model performance and evaluating the arithmetic capabilities of transformers.
- Presented state-of-the-art NLP research articles to onepoint data scientists on a monthly basis.

January–October 2020 **Research Intern - onepoint**, *Paris, France*

- Developed state-of-the-art models to evaluate their capabilities in industrial setting.

November–January 2020 **Data Science Advisor - GMS consulting**, *Rabat, Maroc*

Design of a one-week acculturation training in Data Science for the Moroccan Ministry of Public Transport. Implementation of the training for non-expert employees of the Ministry.

March–August 2019 **Final year internship in data science - Weave**, *Saint-Lazare, Paris*

Analyzed methods to efficiently retrieve answers from a database of question-answer pairs. Compared performance of :

- BERT for semantic similarity.
- BERT question answering.
- Rasa for conversational agents.

June–September 2018 **Trainee developer - EI-Technologies**, *Levallois-Perret, Paris*

- Assignment for **MOSS SAS** : Generic and automated XML schema translation between companies (MOSS SAS and THALES) using XSLT.
- Assignment for **GrasSavoye** : Creation of an interface for customer service management. Implementation of a ChatBot on the interface. Team of 6 people.

Research

2024 **CodeInsight : A Curated Dataset of Practical Coding Solutions from Stack Overflow**, *Findings of Association for Computational Linguistics*, Bangkok

Abstract : Introduces a dataset of 3,409 examples for code generation, featuring clarified intents, code snippets, and related unit tests. Covers libraries such as `Pandas`, `Numpy`, and `Regex`, refined for reduced data contamination. Evaluated on models like Mistral 7B, CodeLLaMa 13B, Starcoder 15B, and GPT-4 to highlight model strengths and weaknesses in coding tasks.

2022 **The impact of lexical and grammatical processing on generating code from natural language**, *Findings of Association for Computational Linguistics*, Dublin

Abstract : Considering the seq2seq architecture of TranX for natural language to code translation, we identify four key components of importance : grammatical constraints, lexical preprocessing, input representations, and copy mechanisms. To study the impact of these components, we use a state-of-the-art architecture that relies on BERT encoder and a grammar-based decoder for which a formalization is provided. The paper highlights the importance of the lexical substitution component in the current natural language to code systems.

Under Submission **RETROcode : Scaling Natural Language to Code Translation with a Syntactic and Database-Driven Approach**

We introduce RETROcode, a model that enhances code generation by utilizing an external code database as an auxiliary resource, reducing the need for larger models or datasets. RETROcode ensures syntactic validity through grammar-constrained generation, making it effective in low-resource settings. Our results show that RETROcode outperforms similarly sized models and even surpasses Codex 12B on development aid datasets.

Under Submission **When Should You Trust the Old Policy ? Investigating the Influence of Pre-Training on Language Model Exploration in Reinforcement Learning**

We explore how pre-training affects the exploration dynamics of language models in reinforcement learning tasks. Focusing on a basic arithmetic task, we propose a modification to the KL divergence penalty that more effectively balances exploration and proximity to the pre-trained model, improving the model's ability to optimize long-term goals.

2023–2024 **Reviewer**, *ACL Rolling Review (ARR)*

Peer-reviewed submissions for the ACL Rolling Review, a continuous review process that hosts conference tracks for major venues like ACL, EMNLP, and NAACL.

Skills

French Native

English Full professional ability

Spanish Proficient

Computing PYTHON, Angular, XML - XSLT, XPATH - LaTeX

Teaching Experience

2022–2024 **Teaching Assistant**, *Université de Paris Cité*, Paris

Assisted in undergraduate courses on Natural Language Processing, Machine Learning, and Software Engineering. Responsibilities included grading, mentoring students, and giving tutorials on Python programming.