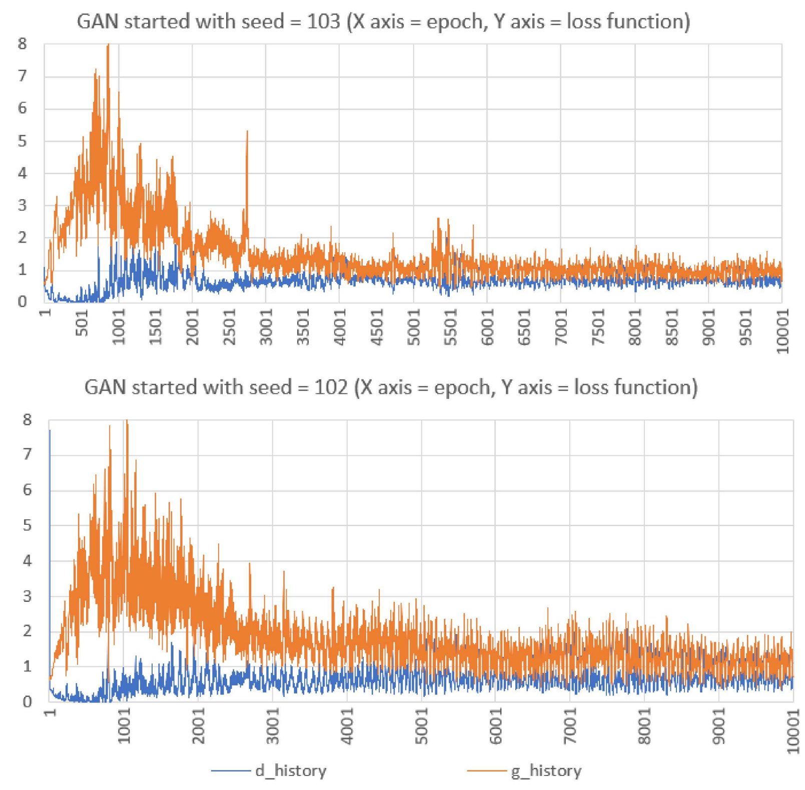
# State-of-the-Art GenAI and LLMs Creative Projects, with Solutions



Vincent Granville, Ph.D. | www.MLTechniques.com | Version 4.3, May 2024

# Preface

With tutorials, enterprise-grade projects with solutions, and real-world case studies, this coursebook covers state-of-the-art material on GenAI, generative adversarial networks (GAN), data synthetization, and much more, in a compact format with many recent references. It includes deep dives into innovative, ground-breaking AI technologies such a xLLMs (extreme LLMs), invented by the author.

The focus is on explainable AI with faster, simpler, high-performance, automated algorithms. For instance: NoGAN, new evaluation metrics, xLLM (self-tuned multi-LLM based on taxonomies with application to clustering and predictive analytics), variable-length embeddings, generating observations outside the training set range, fast probabilistic vector search, or Python-generated SQL queries. I also discuss alternatives to traditional methods, for instance to synthesize geo-spatial data or music.

**Step 1: Summary tables**. Besides embeddings, what are the summary tables and underlying data structure, for each of them? How are they linked to the output returned to a user query? How to adapt the methodology if the input source is Wikipedia rather than Wolfram?

**Step 2: Issues with Python libraries**. By looking at my comments in the code or by experimenting yourself, identify bad side effects from standard Python libraries such as NLTK. Examples include stopwords, auto-correct, and singularize. The problems arise due to the specialization of my xLLM, but is otherwise minor in generic applications targeted to non-experts. Suggest workarounds.

**Step 3: Improving xLLM**. I implemented several enhancements, for instance: ignoring $N$-grams (permutations of $N$ tokens) [Wiki] not found in the crawled data, using normalized PMI in embeddings as the association metric or weight between two tokens, working with variable-length embeddings (VLE) rather than vectors, managing accented characters, not combining tokens separated by punctuation signs, and minimizing stemming such as singular form. Find other potential enhancements and how to implement them. For instance, making sure that "analysis of variance" and "ANOVA" return the same results, or "San Francisco" is not split into two tokens. Hint: use mappings (synonyms) and double-tokens treated as a single token. Treat uppercase and lowercase differently.

Figure 1: The first three steps in the main xLLM project

This textbook is an invaluable resource to instructors and professors teaching AI, or for corporate training. Also, it is useful to prepare for job interviews or to build a robust portfolio. And for hiring managers, there are plenty of original interview questions. The amount of Python code accompanying the solutions is considerable, using a vast array of libraries as well as home-made implementations showing the inner workings and improving existing black-box algorithms. By itself, this book constitutes a solid introduction to Python and scientific programming. The code is also on my GitHub repository.

## About the author

Vincent Granville is a pioneering data scientist and machine learning expert, co-founder of Data Science Central (acquired by TechTarget), founder of MLTechniques.com, former VC-funded executive, author and patent owner.

Vincent's past corporate experience includes Visa, Wells Fargo, eBay, NBC, Microsoft, and CNET. Vincent is also a former post-doc at Cambridge University, and the National Institute of Statistical Sciences (NISS). He published in *Journal of Number Theory*, *Journal of the Royal Statistical Society* (Series B), and *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He is also the author of multiple books, available here. He lives in Washington state, and enjoys doing research on stochastic processes, dynamical systems, experimental math and probabilistic number theory.

# Contents