Léon-v1 : Podcast Research Assistant

Project : TOAQ Research Radio
Date : February 22, 2026
Report version : 1.0

Abstract

This document details the evolution, training, and validation of the **Léon-v1** model. Succeeding the v0 prototype, this Mistral 7B architecture (optimised via Unsloth/QLoRA) resolves narrative redundancy issues through the implementation of temporal positioning. This report analyses the new distillation methodology on 41 documents (450 pairs), the learning dynamics on NVIDIA A100 infrastructure, and demonstrates the model's ability to generate fluid, technically accurate podcast scripts that are natively compatible with speech synthesis (SSML).

Parameter	**Detail**
Base Model	Mistral 7B v0.3 (Unsloth)
Dataset	41 Documents ($\approx$ 450 Pairs)
Architecture	Transformer Decoder-Only (LoRA Adapter)
Hardware	NVIDIA A100 (40GB VRAM)
Licence	Internal Use TOAQ

Author : Côme Bruneteau

Contents

1 Introduction and Context

1.1 The TOAQ Research Radio Project

TOAQ's goal is to democratise access to cutting-edge scientific research by transforming complex academic publications into accessible and engaging audio formats (podcasts). Over the past few months, experiments have revealed several technical challenges related to automation using generalist LLMs (inadequate rhythm, Markdown formatting that is toxic for TTS, repetitions during long generations).

The **Léon-v1** model has been specifically engineered to solve these issues and embody the laboratory's expert and warm radio host, ensuring strict fidelity to the source documents (zero extrapolation).

1.2 The Problem of the Long Generation

Converting a complete scientific publication requires sequential chunking. During the prototyping phase (v0), the model treated each section as a new episode, systematically generating extraneous welcome and conclusion formulas. V1 introduces an elegant solution to this problem through data engineering.

2 Data Engineering

The quality of Léon-v1 is based on a "Knowledge Distillation" strategy using a heavy instructor model (Mistral Large) on a real corpus.

2.1 Dataset Creation Pipeline V1

The final dataset was drastically expanded to reach ≈ **450 pairs** (Input: LaTeX / Output: Script) from 41 research papers. The main innovation lies in the integration of **Temporal Positioning** directly into the training prompts:

- `[Position: START]` : Teaches the model to generate a hook and present the programme.
- `[Position: MIDDLE]` : Forces the model to avoid greetings, favouring varied narrative transitions and a direct dive into technical popularisation.
- `[Position: END]` : Triggers the analytical conclusion and the end-of-episode signature.

This explicit packaging completely eliminates the need for post-processing Python scripts (Regex) to clean up format hallucinations.

3 Legal and Ethical Compliance (Data Compliance)

3.1 Legal Framework: Directive (EU) 2019/790

The training dataset is compiled by extracting open access scientific documents. This approach falls within the European legal framework defined by the Directive on Copyright in the Digital Single Market (DSM Directive), specifically Article 4 relating to the exception for text and data mining (TDM).

3.2 Verification Protocol (Opt-Out)

Before any data is ingested, our pipeline performs an automated check (analysis of `robots.txt` files and explicit CC-BY/CC0 licences). In the absence of explicit reservations (TDM-Reservation protocol), the use of data for fine-tuning is considered legitimate.

4 Training Methodology

Training was performed using the Quantised Low-Rank Adaptation (QLoRA) technique via the Unsloth library, enabling extreme memory optimisation.

4.1 Infrastructure and Performance Hardware

The transition to V1 required greater computing power, deployed on Google Colab Pro:

- **Accelerator :** 1x NVIDIA A100 (40GB VRAM).
- **VRAM efficiency :** Using only **7.6 GB out of the 40.0 GB** available for the GPU, proving the scalability of the quantisation method.
- **System RAM:** Light footprint of **6.1 GB out of 83.5 GB**.
- **Calculation rate:** 5,268 samples per second with a total runtime of 455.59 seconds.

4.2 Hyperparameters and Optimisation

The NVIDIA A100 architecture enabled the adoption of the **Bfloat16** (BF16) format for improved numerical stability, as well as an increase in *batch size*.

Hyperparameter	**Value**
Base Model	Mistral 7B (v0.3)
LoRA Rank (r) / Alpha	16 / 16
Target Modules	`q, k, v, o, gate, up, down_proj`
Learning Rate	1×10^{-4}
Batch Size (per device)	4
Gradient Accumulation	4 steps (Global Batch: 16)
Max Steps	150 ($\approx$ 6 Epochs)
Optimizer	AdamW 8-bit
Warmup Steps	15 (10% of the total)
Precision	BF16 (Native A100)

Table 1: SFTTrainer configuration (Léon-v1)

5 Analysis of Learning Dynamics

The increase in data volume (450 pairs) coupled with learning over 150 steps demonstrates exceptional convergence.

5.1 Training Loss Curve

The model began its learning with a loss of 2.43 (Step 1) and converged smoothly and stably towards 1.06 (Step 150).

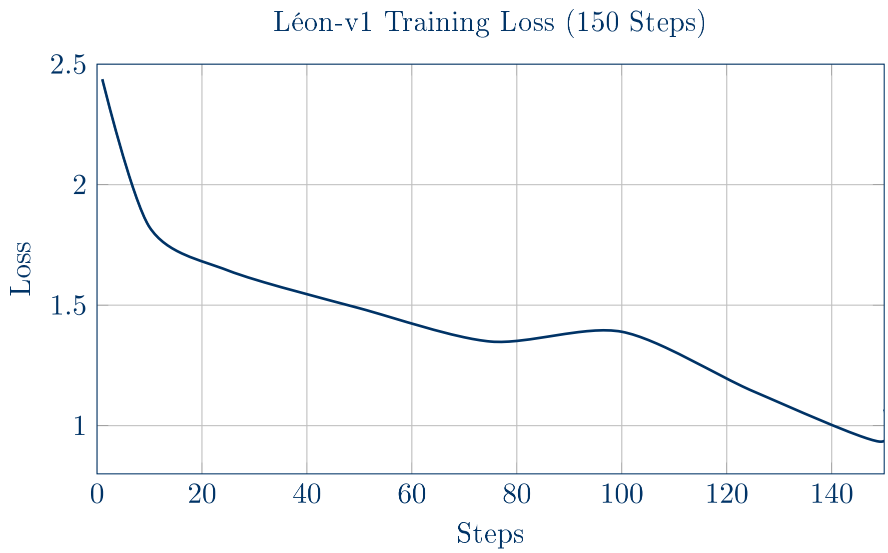

Figure 1: Convergence of V1 training. Global minimum reached at step 149 (0.9341).

Interpretation : The model does not suffer from overfitting (the loss does not collapse towards zero). The equilibrium point around 1.00 indicates that the network has perfectly assimilated the narrative structure (SSML, smooth transitions) while retaining its semantic flexibility to popularise LaTeX.

6 Perspectives and Roadmap

The validation of the Léon-v1 architecture marks a turning point for the automation of TOAQ Research Radio. The short- and medium-term roadmap includes:

1. **Multimodal Textual Integration:** Inclusion of textual descriptions of tables and figures (Vision-to-Text) in the dataset to compensate for the lack of structured data.

2. **100% Autonomous ArXiv Pipeline:** Deployment of the automatic scraper connected to Léon's local GGUF inference, for automated publication.

A Timestamp Markup Demonstration (Léon-v1)

The following excerpt illustrates the model's ability to comply with the `[Position: MIDDLE]` tag constraint. Note the complete absence of greetings, the smooth transition ("What's fascinating here...") and the strategic placement of SSML silences.

```
What's fascinating here is how the very tools we've built to foster trust in AI-checklists, fairness metrics, explainability toolkits-are themselves not neutral. <break time="0.5s" /> They carry the fingerprints of the cultures and power structures that created them. <break time="0.7s" />

Think about it: when UNESCO's 2021 Recommendation on AI Ethics was adopted by 193 states, it was hailed as a global consensus. But consensus doesn't mean universality. <break time="0.4s" /> These frameworks, these procedural safeguards, they're steeped in Western liberal epistemologies. <break time="0.6s" /> And while those values have merit, they're not the only way to conceptualize trust.

Now, here's the tension. <break time="0.3s" /> Replacing entire toolchains isn't just impractical-it's nearly impossible. <break time="0.5s" />

But what if the answer isn't to scrap what we have, but to reimagine how we use it? <break time="0.4s" /> To redesign tools so they're not permanently tethered to Western assumptions, but instead, adaptable-locally appropriated, culturally attuned.
```

Listing 1: Native Generation V1 (Zero Post-Processing)

B Training Metrics

Data	**Value**
total_flos	1.0549846491375206e+17
train/epoch	6
train/global_step	150
train/grad_norm	1.07926
train/learning_rate	0.0
train/loss	1.0676
train_loss	1.39294
train_runtime	455.5993
train_samples_per_second	5.268
train_steps_per_second	0.329v

Table 2: Training metrics