

Amirhossein Yousefiramandi

Applied Scientist · LLMs / NLP / Generative AI

Barcelona, Spain | +34 645 663 255 | [linkedin.com/in/amirhossein-yousefiramandi](https://www.linkedin.com/in/amirhossein-yousefiramandi) | github.com/amirhossein-yousefi | [Google Scholar](#)

SUMMARY

Applied scientist with 6+ years productionizing NLP / LLM research at scale — pretraining domain-specific language models, designing taxonomy-guided classifiers, and shipping research-grade systems on AWS. Currently **Senior Machine Learning Engineer at IQVIA** (Barcelona), owning the production ranking models, data pipeline, evaluation framework, and monitoring behind **clinical-trial site tiering**; previously nearly three years at Clarivate on patent-domain NLP and generative AI. **First-author on four arXiv preprints** (ModernBERT for patents; causal-LLM classification; synthetic-patent-data study; patent-embeddings benchmark) and **sole-author on two 2026 preprints** (Forward-Forward layer free-riding; DRL for deadline-constrained coded caching); three patents (sole inventor on one, first inventor on one, co-inventor on one); two published, one pending USPTO applications. Comfortable owning the loop from hypothesis → controlled experiment → ablation → deployed service.

EXPERIENCE

IQVIA

Senior Machine Learning Engineer

Barcelona, Spain

Jun. 2026 – Present

- Own the ranking models and the data pipeline behind **clinical-trial site tiering** end to end — the system trial teams use to choose investigator sites for studies. The model work is data-centric: improve what the **two-tower ranker** is fed, not the architecture.
- Build and operate the **evaluation framework** that decides whether a candidate model replaces the one in production, ranking each against the incumbent on versioned datasets. Guard the measurement itself, not only the model: drift and **feature-space-collapse detection**, plausibility ceilings on predictions, de-duplication of identical retrains. An evaluation run that reports nothing alerts like a failed one.
- Moved nightly ingestion off cron-offset scheduling onto orchestrated **Airflow** DAGs with real dependencies, retries, data-quality gates and end-to-end failure alerting. **The nightly window fell from ~9 h to ~5 h.**

Clarivate

Senior Data Scientist (Applied Science)

Barcelona, Spain

Aug. 2023 – Jun. 2026

- Designed and shipped the ML core of the **Derwent AI Classifier**, which mapped a customer's patents onto their own taxonomy: hierarchical classifiers trained on demand where labels existed, LLM labeling where they did not. **\$10M+ in incremental revenue** and a **30% lift in renewal contracts**; two published USPTO applications cover the work.
- Replaced a manual examiner workflow with a **multi-agent prior-art and novelty research system** on AWS, with guardrails against prompt injection and data leakage, and built the per-release evaluation suite that made its improvements measurable and customer-reportable. Sole inventor on the patent filing that covers it.
- Pretrained patent-domain masked language models from scratch on a curated **60M+ patent corpus** with a custom tokenizer. The resulting model beat general ModernBERT on three of four internal benchmarks at over three times PatentBERT's inference throughput (arXiv:2509.14926).

Double Avatars

AI Engineer

Spain

Feb. 2023 – Aug. 2023

Virtual try-on and 3D avatarization models for e-commerce.

Freelance

Freelance Data Scientist

Spain

Sep. 2022 – Feb. 2023

Semantic search improvements (MRR@15 **+8%**) and multimodal price-estimation and recommender systems (MAE **−12%**).

Semrush

Data Scientist

Spain

May 2022 – Aug. 2022

Active-learning NLP pipelines on large-scale web text.

CafeBazaar

Data Scientist

Tehran, Iran

Nov. 2020 – May 2022

Semantic app search (Recall@20 **+9%**); POI extraction with YOLO / OCR / SATRN (F1 **53% → 65%**, half the inference time); multi-GPU segmentation training (**+11% mIoU**).

Baharan

Data Scientist

Iran

Feb. 2020 – Oct. 2020

Facial analysis and recognition models for surveillance systems.

- **Yousefiramandi, A.** & Cooney, C. (2026). *When Does Synthetic Patent Data Help? Volume-Fidelity Trade-offs in Low-Resource Multi-Label Classification*. arXiv:2605.24297. **[first author]** — Submitted to ACL Rolling Review; target venue: EMNLP 2026.
- **Yousefiramandi, A.** & Cooney, C. (2026). *Benchmarking Patent Embeddings: A Multi-Task Evaluation of 22 Models Across Retrieval, Classification, and Clustering*. arXiv:2605.24296. **[first author]** — Submitted to ACL Rolling Review; target venue: EMNLP 2026.
- **Yousefiramandi, A.** & Cooney, C. (2025). *Patent Language Model Pretraining with ModernBERT*. arXiv:2509.14926. **[first author, 4 citations]**
- **Yousefiramandi, A.** & Cooney, C. (2025). *Fine-Tuning Causal LLMs for Text Classification: Embedding-Based vs. Instruction-Based Approaches*. arXiv:2512.12677. **[first author]**
- **Yousefiramandi, A.** (2026). *Cumulative-Goodness Free-Riding in Forward-Forward Networks: Real, Repairable, but Not Accuracy-Dominant*. arXiv:2605.06240. **[sole author]** — Submitted to NeurIPS 2026.
- **Yousefiramandi, A.** (2026). *Learning Selective Merge Policies for Deadline-Constrained Coded Caching via Deep Reinforcement Learning*. arXiv:2605.15236. **[sole author]** — Submitted to IEEE ICNP 2026; **40.9%** reduction in broadcast-packet expiration ratio.
- **Patent (pending — in verification):** *Prior Art Search Using Artificial Intelligence Agents*. Camelot UK Bidco Ltd. / Clarivate; Docket No. I03.00450000. **[sole inventor]**
- **Patent (published):** *Training an Artificial Intelligence Classifier Using LLM-Labeled Training Data*. U.S. Application No. 19/029,628 (Pub. US20260212116A1, 23 Jul. 2026); PCT counterpart WO/2026/154406, published 23 Jul. 2026. Docket No. I03.00430000 (Camelot UK Bidco Ltd.). **[first inventor]**
- **Patent (published):** *Hybrid Artificial Intelligence Classifier*. U.S. Application No. 18/590,188 (Pub. US20250272608A1); PCT counterpart WO/2025/181626, published 4 Sep. 2025. *Co-inventor*.

SELECTED OPEN-SOURCE PROJECTS

Text-JEPA Experiments & AF-JEPA | *PyTorch, MiniLM, Qwen 2.5-7B* Open-source

- Two empirical studies on JEPA target design for text representations: semantic-card ablation across 4 seeds (vs. raw embeddings / summaries / shuffled controls) on WikiText-2; **AF-JEPA** swaps the EMA target encoder for a frozen pretrained LLM (Qwen 2.5-7B), trained with smooth-L1 alignment on FineWeb-Edu.
- Block-masking (I-JEPA style) + InfoNCE / cosine alignment; raw-embedding target reaches **0.242 ± 0.016** top-1 retrieval vs. semantic cards **0.183 ± 0.008**. GitHub.

SSM Benchmark: Mamba-2 vs. Transformer vs. Jamba | *PyTorch, CUDA, FineWeb-Edu* Open-source

- Apples-to-apples training and inference benchmark of selective state-space models (**Mamba-2**), a Llama-style **Transformer** (RoPE + SwiGLU), and hybrid **Jamba** (Mamba + attention + MoE) on a ~1.5B-token FineWeb-Edu slice with GPT-2 BPE.
- Evaluated on MQAR, induction copy with extrapolation, and selective copy; measures validation perplexity, throughput, peak memory, and decode latency vs. sequence length. Confirms Mamba’s flat decode latency / KV-cache memory vs. Transformer growth, with Jamba in between. Resumable training (atomic checkpoint of model / optimizer / scheduler / RNG / data cursor) survives Colab’s 12-hour session cap. GitHub.

Multilingual LLM Adaptation via MAD-X Adapter Stacking (Swahili Biomedical) | *HF Adapters, AfroXLMR* Open-source

- Composable adapter stack on a frozen **AfroXLMR-large** backbone: Swahili language adapter + English biomedical domain adapter (MLM-trained) + supervised task adapters, stacked at inference (MAD-X / Pfeiffer).
- Targets medical MCQA (MMLU-ProX-Swahili), Swahili NER (MasakhaNER, BC5CDR-Disease projected), and topic classification; training data from NLLB-200 and OpenWHO. GitHub.

Instruction-Tuned LLM for Text Classification (LoRA/QLoRA) | *TRL, PEFT* Open-source

- Instruction-tunes a decoder-only LLM to output exact label strings; LoRA adapters + 4-bit quantization (QLoRA-style); demo reports **~94% accuracy**. GitHub | Model.

Multi-Modal Hateful Content Classification (CLIP/SigLIP Fusion + MTL) | *CLIP, SigLIP* Open-source

- Late-fusion gated attention + per-label multi-task head on MMHS150K; test **macro-F1 0.569, ROC-AUC (macro) 0.783**. GitHub | Model.

Text+Image VLM Fine-Tuning via JSON Generation | *Qwen2-VL, PaliGemma2* Open-source

- Generative classification on MMHS150K: prompt the VLM to return JSON, parse extracted labels, compute multi-label metrics; Qwen2-VL+LoRA **micro-F1 0.611** (test). GitHub | Adapter.

Whisper-based Multi-Task ASR + Voice Intensity | *PyTorch, Transformers, SageMaker* Open-source

- Joint speech recognition + loudness regression on Whisper encoder; multi-task loss $\mathcal{L} = \mathcal{L}_{\text{ASR_CE}} + \lambda \cdot \text{MSE}$.
Reported: **WER 4.6976%**, intensity **RMSE 0.7334**. GitHub | Model.

Wav2Vec2 Multi-Task: CTC ASR + Intensity | *PyTorch, HF, SageMaker* Open-source
 • Wav2Vec2-base-960h CTC + intensity head with attention-masked mean pooling; validation **WER $\approx 12.9\%$** ; config-first repo. GitHub.

Image-Text Retrieval (CLIP/SigLIP) | *FAISS-ready, custom Trainer* Open-source
 • Contrastive training/eval on Flickr8k/30k: Flickr8k **R@1 90.7%** (Img→Text), **77.06%** (Text→Img); Flickr30k **R@1 92.3%** / **79.0%**. GitHub.

Multi-Label Emotion Classification (GoEmotions) | *Transformers* Open-source
 • 27-emotion multi-label detection with RoBERTa-base; validation threshold tuning; test **micro-F1 0.528 @ thr 0.84, ROC-AUC (micro) 0.952**. GitHub.

Emotion-Aware TTS Style Transfer | *SpeechT5, WavLM, ECAPA, HiFi-GAN* Open-source
 • StyleAdaptor + StyleSpeakerFusion: fuse WavLM emotion/prosody with ECAPA speaker identity for expressive TTS; Gradio UI. GitHub.

TECHNICAL SKILLS

Research / Modeling: Transformer architectures (MHA, GQA, MoE), attention variants (Flash, paged), positional encodings (RoPE, ALiBi), state-space models (Mamba, Mamba-2, Jamba), RLHF / DPO / GRPO, knowledge distillation, continued pretraining, tokenizer extension, controlled ablation design

Frameworks: PyTorch, Hugging Face (Transformers, TRL, PEFT, Diffusers), MosaicML Composer, MLflow, Scikit-learn, LightGBM

LLMs & Agents: UnSloth (Llama / Gemma), LoRA / QLoRA, DSPy, DeepAgents, LangGraph, LlamaIndex, MCP, Strands Agents, AgentCore, Google ADK, DeepEval

Cloud: AWS (SageMaker, Glue, Lambda, Fargate, Bedrock, Step Functions), Azure AI, Google Gemini integrations

Languages & Tools: Python, C++, Bash, SQL, Git, Docker, ONNX, Jenkins, L^AT_EX

Quantitative Research: Mathematical optimization, mathematical modeling, statistical learning, information theory

Spoken Languages: English (fluent), Spanish (working), Persian (native)

EDUCATION

Sharif University of Technology	Tehran, Iran
<i>M.Sc. in Electrical Engineering (AI)</i>	2018 – 2021
Isfahan University of Technology	Isfahan, Iran
<i>B.Sc. in Electrical Engineering</i>	2014 – 2018

HONORS AND AWARDS

- Promoted at Clarivate within less than a year for exceptional performance.
- Twice promoted at CafeBazaar within one year.
- Ranked 16th nationwide in Iran's master's-program entrance examination.