Thomas Compton CV

LinkedIn   Medium   GitHub

Statement

Senior Data Scientist / ML Engineer specialising in NLP, RAG pipelines, and scalable LLM workflows. Experienced in fine-tuning and evaluating topic models, embedding search, and automating OCR/data preparation. Built and deployed agentic conversational systems (Dialogflow) and multi-agent debate simulators. Skilled in Python, PyTorch, Hugging Face, FAISS, LangChain, AWS, and Docker, with strong academic background and applied open-source contributions.

Experience

School for Business and Society, University of York

Graduate Teaching Assistant & Research Assistant (May 2023 – Aug 2025)
- Designed and tested AI-enabled chatbot to support staff with navigating EDI policy changes.
- Conducted 10 practitioner interviews to inform research on adoption of intelligent systems.
- Delivered seminars and coding workshops, consistently earning strong feedback.

Many Flavours of Wellingborough CIC

Founder & Director (Sept 2023 – Sept 2024)
- Founded, registered, and managed CIC including safeguarding, EDI, and data protection policies.
- Coordinated volunteers and ran a youth drama group with 11 regular students.
- Secured multiple funding awards totalling over £2,700.
- **Mentored** volunteers and student participants, supporting skill development and inclusion.

Projects

Database Debate Simulator (Agentic AI)
- Built a RAG-based system to simulate multi-agent debates using short LLM-driven conversations.
- Demonstrated integration of embeddings, retrieval, and generative AI for interactive reasoning.
- **Skills:** LangChain, FAISS, Hugging Face, Python.

OCR Accuracy Evaluation: GitHub
- Benchmarked WER & CER across EasyOCR, PaddleOCR, Tesseract and LLM-assisted OCR.
- Developed batched workflow with backoff for high-volume OCR jobs.
- **Skills:** EasyOCR, PaddleOCR, Tesseract, Google Gemini, Jiwer, Deepseek, Qwen.

BERTopic Evaluation and Fine-Tuning: GitHub
- Developed stability metrics and evaluation pipeline for repeated BERTopic runs.
- Automated batch selection of best topic models for large-scale corpora.
- **Skills:** BERTopic, Sentence Transformers, scipy, matplotlib, UMAP.

Education

University of York (2021–2025 expected)
- ESRC-funded PhD applying NLP methods to large historical corpora.
- Thesis: *Historical (dis)Continuities in Community Unionism Between The Boot and Shoe Union and Unite Community.*

University of Birmingham (2016–2020)
- MSc (Distinction), International Relations.
- BA (2:1), English and Creative Writing.

Skills

- **LLMs & Agentic AI:** Dialogflow, Google Gemini, OpenAI, Qwen, TrOCR
- **RAG & Vector Search:** LangChain, FAISS, Hugging Face
- **Programming:** Python (3+ years), PyTorch, scikit-learn
- **NLP & ML:** spaCy, NLTK, BERTopic, LDA, Sentence Transformers
- **Data & Tools:** AWS, Docker, GitHub, Gradio, pandas, matplotlib, requests

Open Source Projects

Literature Review: GitHub
- Modular RAG tools for literature reviews using embeddings and FAISS.
- **Skills:** RAG, FAISS, Sentence Transformers, spaCy, scikit-learn.

PDF-to-Speech Reader: GitHub
- Converts PDFs into audio with a repeatable text-to-speech workflow.
- **Skills:** pypdf, pyttsx3, pydub, AWS.

Hansard OCR Dataset: GitHub
- Prepared UK parliamentary debates for OCR training and evaluation.
- **Skills:** TrOCR, torch, BeautifulSoup, transformers.

Samuel Smiles DA: GitHub
- Repo for historical text analysis with embeddings and topic modelling.
- **Skills:** sentence-transformers, FAISS, BERTopic, spaCy, NLTK.

ngram-2-topic: GitHub
- Combines n-gram frequency with embeddings for interpretable topics.
- **Skills:** sentence-transformers, spaCy.

Publications

Articles

Compton, T. (2025a). Beyond the black box: Integrating lexical and semantic methods in quantitative discourse analysis with bertopic. https://arxiv.org/abs/2508.19099

Compton, T. (2025b). Holistic evaluations of topic models. https://arxiv.org/abs/2507.23364

Conferences
- BUIRA Conference (accepted), 2024, *Northampton Boot & Shoe Then & Now*
- BSA WES Conference (accepted), 2023, poster *Northampton Boot & Shoe Then & Now*