Smart Lead Gen: A Generative AI for Hyper-Personalized Professional Outreach

- Milestone 3: Model Architecture.
- Team: Group 7
- Course: DS and AI Lab
- Submission date: October 17th, 2025

The SmartLeadGen project is designed to address a critical inefficiency in professional communication: the trade-off between scalable but impersonal mass outreach and effective but time-consuming manual personalization. To solve this, the project employs a sophisticated, multi-stage model architecture that transforms a simple user request into a hyper-personalized, context-aware, and strategically tailored message. This pipeline is composed of three sequential modules, each optimized for a specific task: Stage 1 handles data collection and structuring, Stage 2 enriches this data with psychographic insights, and Stage 3 synthesizes all available information to generate the final, high-impact output. This modular design ensures that each step of the personalization process is handled by a specialized, efficient, and robust component, culminating in a system that enables high-quality outreach at scale.

Stage 1: Data Ingestion & Parsing Module

The Data Ingestion and Parsing Module serves as the foundational entry point for the entire SmartLeadGen pipeline. Its primary purpose is to gather, clean, and structure publicly available data about a target professional, creating a comprehensive profile that enables deeply personalized communication. The process begins when a user provides two key inputs: the LinkedIn profile URL of the target individual and a clear objective or prompt, such as "Send a job offer for a Software Engineering role".

The module’s technical workflow is methodical and robust. First, it validates the input URL using regular expressions or parsing libraries to ensure it is correctly formatted and accessible. Once validated, the system identifies and extracts data from multiple sources. It scrapes the target's public LinkedIn profile for key information like their name, headline, experience, and skills using tools like Playwright. To gather broader context, it also accesses company websites for product details, news websites for the latest company updates via APIs like NewsAPI, and platforms like GitHub or Medium for technical contributions and blog posts.

Data extraction is performed ethically and responsibly by adhering to robots.txt compliance and implementing rate limiting to avoid overloading servers. A proxy rotation service is used to distribute requests and ensure stability. Following extraction, the raw data undergoes a rigorous cleaning process to remove unwanted characters and formatting, utilizing libraries like re and nltk. The clean data is then organized into a predefined JSON schema with fields for name, headline, career history, skills, recent activity, and company updates . This structured JSON object is the module's final output, serving as a comprehensive input data object for the downstream personality classification and message generation processes.

Stage 2: Personality Classification Module

The Personality Classification Module represents a critical enrichment phase in the SmartLeadGen pipeline, adding a layer of deep psychographic personalization. Its primary objective is to infer a prospect's MBTI personality type by analyzing their publicly available written text, such as recent posts or articles. This inferred personality type serves as a vital input for the final generative model, allowing it to precisely tailor the outreach message's tone, style, and content to resonate most effectively with the recipient's psychological profile.

For this sophisticated text classification task, the chosen architecture is a fine-tuned Transformer-based model, specifically DistilBERT. This model is significantly smaller and faster than its larger counterparts like BERT-base, making it more efficient for deployment without sacrificing the state-of-the-art performance of Transformer architectures. The model consists of an embedding layer, six Transformer blocks that use multi-head self-attention to capture contextual nuances in text, and a final classification head with 16 output neurons corresponding to the 16 MBTI types. By leveraging a pre-trained DistilBERT model, the project utilizes transfer learning to achieve high accuracy with reduced training time and data requirements.

The data pipeline for this module is twofold. For training, the Myers-Briggs and Pandora (Big 5) datasets are merged into a unified corpus, with Big Five labels reparametrized to the MBTI framework. This data is cleaned, tokenized, and used to fine-tune the DistilBERT model. The live inference pipeline, however, executes in real-time within the application. It receives the structured JSON from Stage 1, extracts relevant text from fields like recent_activity, and applies the exact same cleaning and tokenization functions used during training to ensure consistency. The processed text is then fed into the fine-tuned classifier, which predicts an MBTI type. This prediction is added as a new field to the JSON object, creating an "enriched" profile that is subsequently passed to Stage 3.

Stage 3: Fine-Tuned Generative Model

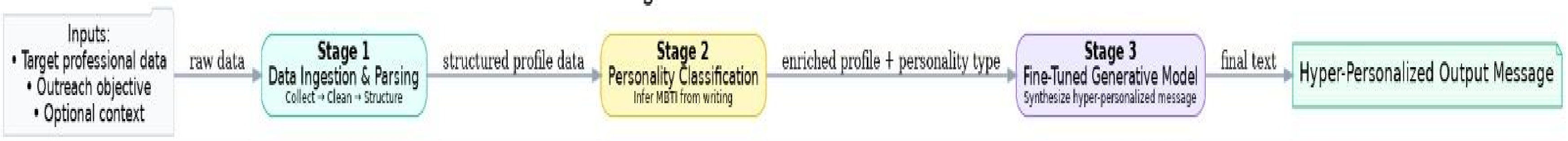
The Fine-Tuned Generative Model is the culminating stage of the SmartLeadGen pipeline, acting as the synthesis layer where data is transformed into communication. This module's purpose is to synthesize all upstream data—the structured professional profile from Stage 1 and the inferred MBTI personality type from Stage 2—to generate a hyper-personalized and context-aware outreach message. The goal is to produce text that is not only professionally coherent but also strategically crafted to maximize engagement and conversion based on the recipient's unique professional and psychological profile.

The core of this stage is the Mistral 7B model, a powerful 7-billion parameter Large Language Model, which has been fine-tuned using QLoRA (Quantized Low-Rank Adaptation). Mistral 7B was chosen over competitors like Llama 3 8B for its superior fine-tuning adaptability, greater computational efficiency, and permissive Apache 2.0 license for commercial use. QLoRA is a parameter-efficient fine-tuning (PEFT) technique that dramatically reduces the computational resources required for training. By quantizing the base model's weights to 4-bit precision and training only small, low-rank adapters, QLoRA reduces the GPU memory requirement from over 60GB for full fine-tuning to just 10-12GB. This approach prevents catastrophic forgetting, converges 2-3x faster, and achieves 97-99% of the performance of full fine-tuning, making it the optimal choice.

The model was trained on the Enron Email Corpus, a dataset of approximately 150,000 filtered corporate emails that provide authentic examples of professional communication patterns. The training data was structured in an instruction-following format, where synthetic professional profiles were reverse-engineered from the emails to teach the model the explicit relationship between a profile, an objective, and a corresponding message. Critically, the prompts are enriched with personality-aware instructions, guiding the model to tailor its tone and style to the recipient's MBTI type—for instance, using concise, data-driven language for an INTJ or a warm, enthusiastic tone for an ENFP. The final inference pipeline dynamically constructs these prompts, generates the message, and performs post-processing checks to ensure quality before delivering the polished draft to the user.

The diagram on the next page depicts the proposed model architecture for the SmartLeadGen project. Link to a clearer image here due to image dimensions mismatch.

High-Level Model Architecture


Inputs:
• Target professional data
• Outreach objective
• Optional context
raw data
Stage 1
Data Ingestion & Parsing
Collect → Clean → Structure
structured profile data
Stage 2
Personality Classification
Infer MBTI from writing
enriched profile + personality type
Stage 3
Fine-Tuned Generative Model
Synthesize hyper-personalized message
final text
Hyper-Personalized Output Message