Round 2

AI/ML Take-Home Task - 48h

Task Overview:

Using the provided sample data, build a mini-pipeline that:

1. **Transcribes** sample_data/sample_audio.wav
2. **Extracts** the top 5 discussion topics from that transcript
3. **Vectorizes** those topics (one-hot, TF-IDF, or embeddings)
4. **Loads** two synthetic user profiles from sample_data/synthetic_users.json (each with a psychometric vector)
5. **Combines** each user's psychometric vector with the topic-interest vector (e.g. concatenation or weighted sum)
6. **Computes** a **compatibility score** between the two users by cosine similarity over their cmobined vectors, and returns an interpretation of that score.
7. Wrap your **pipeline into a FastAPI** application that exposes the following endpoints: POST/transcribe (returns a transcript), POST/summarise (returns topics and summary) and POST/match (returns compatibility score).
8. Use **FastAPI's built-in Swagger/OpenAPI UI** so endpoints can be tested via browser.

Sample Data (in sample_data/):

Audio: 'Is Mars the Future of Humanity?' - Joe Rogan Asks Elon Musk

JSON:

Traits = [openness, conscientiousness, extraversion, agreeableness, neuroticism]

```
[
  { "id": "user_1", "psychometrics": [0.8, 0.4, 0.7, 0.2, 0.9] },
  { "id": "user_2", "psychometrics": [0.3, 0.9, 0.1, 0.6, 0.4] }
]
```

Tech Stack:

Use any combination of:
Transcription: OpenAI Whisper, Hugging Face whisper, AssemblyAI
Topic Extraction: spaCy, Gensim, Transformers (BERT/DistilBERT), keyBERT
Vectorization: scikit-learn TF-IDF, SentenceTransformers, et
Matching Logic: cosine similarity on combined vectors (numpy / scikit-learn)
LLM Fine Tuning: GPT-2, LLaMA via Hugging Face transformers + accelerate or OpenAI API
Orchestration: PyTorch/TensorFlow; LangChain (optional)

Deliverables:

1. GitHub repo with clean, dated, well-documented code.

2. README covering:
   - How to load and use the sample data
   - Architecture & design decisions
   - Topic vectorization and fusion with psychometrics
   - Matching logic and edge-case handling (thresholds or re-sampling)
   - Instructions to run each step

3. Short write-up (≤ 300 words) on:
   - Next steps & improvements
   - Trade-offs made
   - Any creative extensions