Hello Ray, this is my week 2 progress:

Mainly this week has been to get to further know things at a deeper level before we start with the implementation. For this, I went through the exhaustive material and youtube videos you shared.

1. I started with studying the NotebookLM's output for Prof. Neil's persona, here's what I came up with:
   - Firstly, it has mainly discussed about the professor's background, research contributions, awards received, his core principles and values, experience, his approach towards solving a particular problem, frameworks developed by him and also the technical knowledge that he might have (I think we will need more of this during the training process as the accuracy is more essential as compared to replicating the persona , which makes me realize that we will need to focus more on the technical knowledge of the professor and not dump the full information, but present it in a simplified way).
   - While it notes Morgan's communication style, the doc doesn't include enough *direct quotes, phrasing, or typical classroom/consulting examples*.
   - For training purposes, we will need more conversational exmples like e.g., 'He often tells students: 'You can't be great at everything — pick two majors, one minor.'

2. Here's how I think about using the shared YouTube videos, train the digital twin by using the *hybrid approach* -first extract transcripts from videos using high-quality speech-to-text (Whisper, AssemblyAI, etc.) then clean and annotate them by marking catchphrases, stylistic elements, tone shifts, now use these transcripts to train the LLM twin's text reasoning. Additionally, use video to train a voice/visual avatar layer.

3. The ChatGPT's deep research document is really good! It very accurately captures the common phrases, his mindset (logical reasoning) and thinking style tone, etc. other than the technical information. This will really helpful to us for training.

4. Taking inspiration from the paper: **Creativity Has Left the Chat,** I have come up with an approach that will enable us to preserve the standards, tone and mannerisms of the professor without hampering the creativity of the answers:
   This approach is simple: first you brainstorm, then you polish. Start by using a creative "base" model to generate a few very different drafts of the answer—each one should take a different angle and use different examples. Pick the one you like (or combine the best parts), then paste it into a safer, instruction-following "aligned" model to rewrite it in Professor Neil Morgan's voice: clear, evidence-led, structured, with practical steps and a short checklist or questions at the end. You only share

5. Benefits of distillation:
6. Distillation Pipeline:
   - Ingest: Gathering the data: documents, videos, etc
   - Extract: Gist pass, Dense pass, Structure pass
   - Normalize: Using consistent field names across the cards, ensuring overall consistency
   - De-duplicate & Merge: Cluster similar content; keep the clearest version
   - Compress: Turn long prose into short steps, rules, and examples
   - Validate**:** Spot-check with "Would Neil say this?" and "Is this actionable?
   - Export: Save as JSON/CSV + a printable PDF "cheat sheet" for the next training step

Ingest → Extract → Normalize → De-duplicate → Compress → Validate