For our current transcription.py we are using a hugging face transformer and using the model s2t-small-librispeech-asr.
Now what would be the best method to formatting the transcribed text,

Using an AI Model:

1. GPT-4 (OpenAI API) – Can restructure text into paragraphs, outlines, or bullet points.
    - Pricing: GPT-4 API costs approximately $0.03 per 1,000 input tokens and $0.06 per 1,000 output tokens.
2. T5 (Text-to-Text Transfer Transformer) – Fine-tune for text summarization and reformatting.
    - Pricing: Free if running locally; cloud-based solutions like Google Cloud AI may have usage costs.
3. BART (Facebook AI) – Great for cleaning up text and summarizing.
    - Pricing: Free to use with Hugging Face models locally; cloud-based deployments may incur costs.

Integrating using our current code:

Use **transformers pipeline** for text summarization and formatting:

1. print(formatted_text)
    - Pricing: Free if using Hugging Face models locally; Inference API starts at $0.02 per 1,000 characters.
2. Using a Fine-Tuned Model
    - Train a model on well-formatted transcripts.
    - Use datasets from Hugging Face to collect structured speech-to-text data.
    - Pricing: Free if training on local hardware; GPU cloud services (AWS, GCP, Azure) may cost $0.50 - $5 per hour.
3. Post-Processing with Python NLP:
    - Sentence Segmentation: nltk.sent_tokenize(transcribed_text) (Free)
    - Grammar Correction: gramformer or GingerIt
        - GingerIt Premium starts at $13.99/month.
    - Regex Cleanup: Removing unwanted symbols or fillers (Free).