1) How would you handle ambiguous or missing medical data in the transcript?

Sol: When dealing with ambiguous or missing medical data in a transcript, I'll take a structured approach:

1. Use Context Clues If a symptom or diagnosis isn't explicitly mentioned but is implied, I'd infer based on the surrounding information.
2. Default to "Unknown" or "Not Provided" instead of assuming details; I'd leave placeholders for a physician to review.
3. Leverage AI Models for Suggestions NLP models can detect missing details by analysing similar cases.
4. Prompt for Clarification If it's an interactive system, I'd ask follow-up questions to fill in gaps.

2) What pre-trained NLP models would you use for medical summarisation?

Sol: For medical summarisation, I'd use:

1. BioBART – A biomedical variant of BART, fine-tuned for medical text summarization.
2. PubMedBERT – Trained in biomedical literature, useful for extracting key details from medical transcripts.
3. ClinicalT5 – A T5-based model fine-tuned for clinical text summarization.
4. BERTSUM – A summarization model built on BERT, effective for extracting important sentences.

3) How would you fine-tune BERT for medical-sentiment detection?

Sol: To fine-tune BERT for medical sentiment detection:

1. Dataset Preparation – Collect labeled medical dialogues
2. Tokenization – Use a pre-trained BERT tokenizer to convert text into tokens.
3. Model Fine-Tuning – Use BERTForSequenceClassification from Hugging Face, adding a classification head.
4. Training – Fine-tune on labeled medical data using cross-entropy loss and AdamW optimizer.
5. Evaluation – Test using F1-score, accuracy, and validate on real-world medical transcripts.

4) What datasets would you use for training a healthcare-specific sentiment model?

Sol: For training a healthcare-specific sentiment model, you can use these datasets:

1. MIMIC-III / MIMIC-IV – Large ICU patient records with clinical notes (requires access approval).
2. i2b2 NLP Challenge Datasets – Annotated clinical text for various NLP tasks.
3. MedDialog – Medical conversations between doctors and patients.
4. PubMed & ClinicalTrials.gov – Can be scraped for medical text with sentiment annotation.

5) How would you train an NLP model to map medical transcripts into SOAP format?

Sol: To train an NLP model for mapping medical transcripts into SOAP format, you first need a well-structured dataset, such as MIMIC-III or i2b2, that contains both unstructured transcripts and corresponding SOAP notes. The next step is preprocessing, where transcripts are cleaned, tokenized, and annotated with SOAP sections. A transformer-based model like BERT, T5, or GPT can then be fine-tuned for text classification and summarisation. The training process involves feeding annotated data into the model to learn the structure and context of SOAP notes. Finally, the model is evaluated using metrics like BLEU, ROUGE, and F1 to ensure accuracy.

6) What rule-based or deep-learning techniques would improve the accuracy of SOAP note generation?

Sol: Improving the accuracy of SOAP note generation can be achieved through both rule-based and deep-learning techniques. Rule-based methods use medical ontologies like UMLS to identify symptoms, diagnoses, and treatments, ensuring structured data extraction. Regular expressions and dependency parsing with spaCy can help refine entity recognition. Deep learning approaches, such as fine-tuning transformer models like T5 or BART, improve summarization accuracy by learning contextual patterns in medical transcripts. Combining these techniques in a hybrid model—where rule-based methods handle structured data and deep learning refines contextual understanding—ensures higher accuracy and coherence in generating SOAP notes.