Task 2: Deliverable — LLM Prompting Strategy

Chosen Model

Model: GPT-OSS-120B (High)
Provider: OpenAI (hosted on Amazon Bedrock)
Context Window: 131,000 tokens
Artificial Analysis Intelligence Index: 58
Cost: ~$0.26 per 1M tokens
Median Speed: 310 tokens/s
Median First-Chunk Latency: 0.45 s

Rationale for Choice

- **Large Context Window (131k):** Fits long meeting transcripts (up to 1–2 hours) without truncation.
- **Strong Comprehension (AAII 58):** Excellent at reasoning, summarization, and contextual expansion.
- **Low Cost:** Only $0.26 / 1 M tokens — highly economical for large-scale usage.
- **Fast Response:** Suitable for interactive user experience through AWS Lambda.
- **AWS Integration:** Fully compatible with Amazon Bedrock → easy to call from Lambda and integrate with S3 + DynamoDB.

Prompt Template: You are an AI meeting assistant that expands concise bullet points into detailed, structured notes.
Context: {{meeting_transcript}}
User Request:
Expand the following concise note into a detailed note: "{{concise_note}}"
Requirements:

- Include all relevant details from the meeting related to this point.
- Maintain factual accuracy and original speaker context.
- Write clearly and professionally, in paragraph form.
- Avoid repetition or unrelated information

Key Elements of the Prompt

Element	Purpose
System Role Instruction	Guides model tone and objective (assistant for meeting documentation).

Structured Input Blocks	Separates transcript and user note for clarity.
Explicit Output Rules	Ensures consistency, coherence, and professional tone.
Expansion Directive	Focuses model output on detailed elaboration, not generic summaries.

Expected Output

A coherent, well-structured paragraph (or multiple paragraphs) expanding on the original concise note, preserving the factual context and tone of the meeting — suitable for embedding directly under the bullet point in the user interface.