```
3. **Natural follow-ups**: Questions should sound like logical next questions about mentioned topics
4. **Challenging abstention**: Make it harder for LLM to recognize information is missing
5. **Domain consistency**: Questions should fit naturally within the plan's subject matter

Generate 8-15 abstention questions that test the LLM's ability to appropriately say "Based on the provided
     chat, there is no information related to [topic]" when information is missing or unavailable in the plan
     .

NOTE: Only output the JSON array without any explanation before or after the list.
```

Listing 14: Abstention probing question generation prompt

```
You are tasked with generating a probing question to test preference following capabilities of LLMs. You will
     be given a bullet point containing a user preference and corresponding conversation turns where this
     preference was mentioned.

Your task is to create ONE question that tests whether an LLM remembers and respects the user's stated
     preference when making recommendations or providing advice.

## INPUT DATA
- **BULLET POINT**: <bullet_point>
- **CONVERSATION TURNS**: <conversation_turns>

## CRITICAL REQUIREMENT: COMPLETELY NEUTRAL QUESTION
- The question MUST BE COMPLETELY NEUTRAL with NO hints about any preference
- The question MUST NOT contain ANY words that suggest evaluation criteria
- The question should be so neutral it could test ANY type of preference in that domain
- The LLM should only know about the preference from previous conversation history

## MANDATORY PREFERENCE ANALYSIS STEP
BEFORE writing the question, you MUST:
1. **Extract ALL preference-related words** from the bullet point
2. **List ALL forbidden terms** including synonyms and related concepts
3. **Verify your question contains NONE of these terms**

## FORBIDDEN QUESTION ELEMENTS
[Examples]

## QUESTION LANGUAGE REQUIREMENTS
- Write questions as if the USER is asking them naturally
- Use first person ("I", "my", "me") when referring to the user
- Use second person ("you") when addressing the assistant
- Avoid phrases like "according to the conversation", "based on what was discussed", "from our chat history"
- Make questions sound conversational and natural
- Questions should flow naturally as if continuing the conversation
- NEVER mention the preference, decision criteria, or reasoning from the bullet point

## CHAT ID TRACKING REQUIREMENT
- You MUST identify which specific chat_id(s) contain the preference information
- List ALL chat_ids where the preference was mentioned or demonstrated
- NOTE: If the answer is spread out between multiple chat_ids, group them in one list
- NOTE: DO NOT INCLUDE chat_ids in the answer
- Use the exact chat_id numbers from the conversation turns

## OUTPUT FORMAT
Return your analysis in this exact JSON format:
{
    "question": "", "preference_being_tested": "", "expected_compliance": "", "compliance_indicators": [], "
          non_compliance_signs": [],
    "difficulty": "medium", "preference_type": "","source_chat_ids": []
}

## IMPORTANT REQUIREMENTS
1. **Preference-triggering question**: Question must create a situation where the stated preference should
     guide the response
2. **Clear compliance expectations**: Define what respecting the preference looks like
3. **Measurable indicators**: Provide specific signs of following vs. ignoring the preference
4. **Natural question phrasing**: Question should sound realistic and conversational
5. **Preference relevance**: Question must relate to the same domain/context as the stated preference

Generate ONE preference following question that tests whether the LLM remembers and applies the user's stated
     preference when providing recommendations or advice.

NOTE: Only output the JSON object without any explanation before or after.
```

Listing 15: Preference following probing question generation prompt

```
You are tasked with generating a probing question to test event ordering capabilities of LLMs. You will be
     given multiple related bullet points about the same topic/theme and the corresponding multi-turn dialogs
      between a user and assistant that incorporate these mentions across different conversation sessions.

Your task is to create ONE question that tests whether an LLM can recall the chronological order in which
     topics were MENTIONED in the conversation, regardless of when the actual events occurred in real life.

## INPUT DATA
- **BULLET POINTS**: <bullet_points>
- **CONVERSATION TURNS**: <conversation_turns>

## CRITICAL REQUIREMENTS: NO SPOILERS OR TIME HINTS
- The question MUST NOT list, mention, or hint at the specific events/mentions being tested
```