Three significant practical examples of incorrect responses from large language models (**LLMs**) in tasks requiring **abductive reasoning** (inferring the best explanation for observations) are outlined below.

1. **Temporal ordering in commonsense scenarios**

   Prompt: "David arrived after Joe. Joe arrived before me. John arrived after David. Who arrived first?"

   Correct abductive inference: Joe (chain: Joe → David → John, with Joe before the speaker).

   LLM failure (observed in ChatGPT variants): Models often provide incorrect answers (e.g., "John" or "me") or fail to resolve the order, due to weak backward causal chaining from effects to prior causes.

2. **Psychological reasoning and theory of mind**

   Prompt: Classic false-belief task variants (e.g., Sally places a ball in a basket, leaves; Anne moves it to a box. Where will Sally look upon return?).

   Correct abductive inference: The basket (Sally's outdated belief).

   LLM failure (documented in ChatGPT evaluations): Models sometimes predict the box, failing to abduce the character's false belief and instead reasoning from current reality or training patterns.

3. **Clinical diagnosis from symptoms**

   Prompt: Hypothetical patient symptoms requiring inference of underlying cause (e.g., in mARC-QA benchmark cases).

   Correct abductive inference: Prioritize the most plausible disease fitting all observations, including rare conditions.

   LLM failure (in models like o1, Gemini, Claude): High error rates in abductive steps, often defaulting to common diagnoses or missing key explanatory links, performing poorly relative to human physicians.