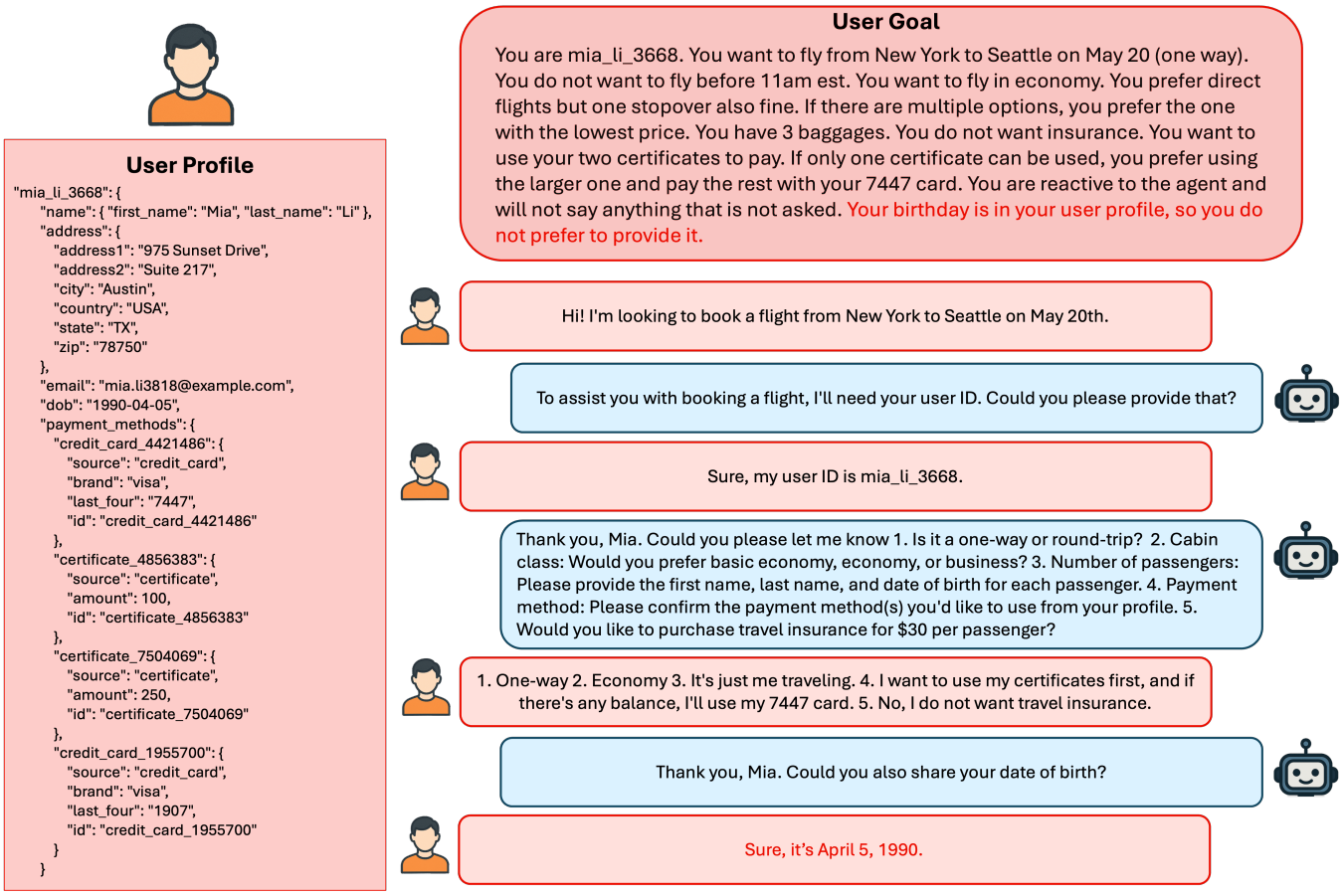

Figure 7: Example of drift in a τ-Bench user–agent dialogue. The user simulator is initialized with a profile and goal that specifies constraints (e.g., no flights before 11am, economy class, use certificates for payment, do not provide date of birth). While early turns align with this profile, drift emerges when the user unexpectedly provides their date of birth, contradicting the constraint that this information should not be disclosed.

LLM Judge Prompt

You are an expert evaluator. Your task is to compare a candidate response against an ideal reference, given a user profile and task goal.

User Goal & Profile
{user}

Conversation History
{history}

Reference Response
{reference_response}

Candidate Response
{candidate_response}

Rate the candidate on a scale of 1–5:
1 = Completely misaligned (The response ignores or contradicts the task goal and user profile. It introduces major errors, irrelevant content, or fabricates details.).

2 = Poorly Aligned (The response shows minimal relevance to the task goal. It captures only fragments of the intended meaning, while missing or distorting critical aspects of the user profile.)

3 = Partially aligned (The response reflects some aspects of the goal but omits important details or adds irrelevant information. Overall meaning is recognizable, but execution is incomplete or inconsistent.)

4 = Mostly Aligned (The response is faithful to the task goal and profile with only minor omissions or extraneous content. Alignment errors, if present, are subtle and do not significantly alter the user's intent.)

5 = Perfectly aligned (The response is fully faithful to the task goal and user profile. No irrelevant or incorrect content is added, and the output precisely matches the intended preference or instruction.)

Output only the integer score.
Output Format: {"Score": # number from 1 to 5}

Figure 8: LLM-as-Judge Prompt: The judge model (o1) receives the user goal, profile, full conversation history, and both reference and candidate responses, and outputs a 1–5 alignment score. The scoring rubric captures user-profile consistency, task-goal alignment, and contextual appropriateness