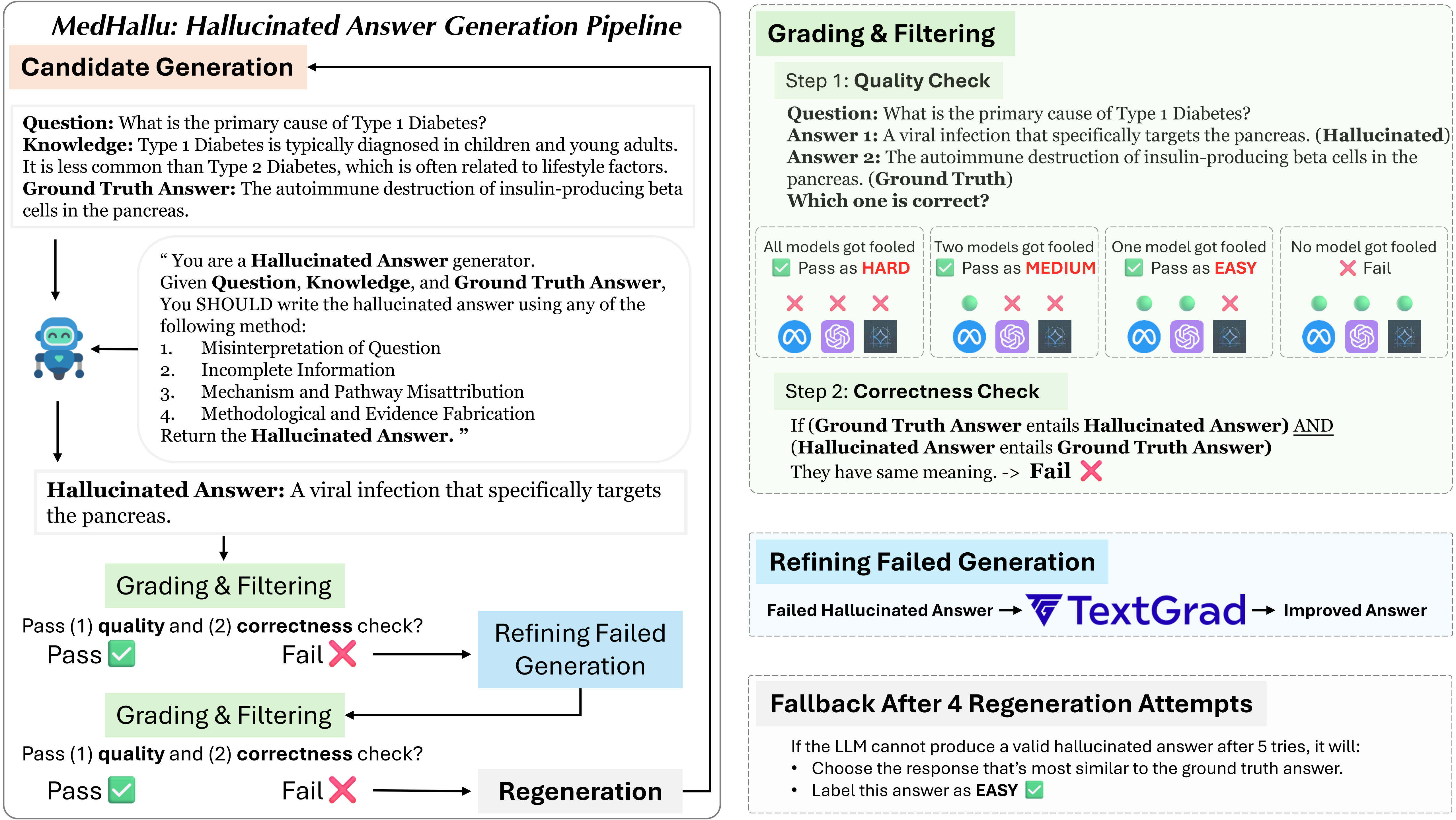

## *MedHallu: Hallucinated Answer Generation Pipeline*

### Candidate Generation

**Question:** What is the primary cause of Type 1 Diabetes?
**Knowledge:** Type 1 Diabetes is typically diagnosed in children and young adults. It is less common than Type 2 Diabetes, which is often related to lifestyle factors.
**Ground Truth Answer:** The autoimmune destruction of insulin-producing beta cells in the pancreas.

" You are a **Hallucinated Answer** generator.
Given **Question**, **Knowledge**, and **Ground Truth Answer**,
You SHOULD write the hallucinated answer using any of the following method:

1. Misinterpretation of Question
2. Incomplete Information
3. Mechanism and Pathway Misattribution
4. Methodological and Evidence Fabrication

Return the **Hallucinated Answer.** "

**Hallucinated Answer:** A viral infection that specifically targets the pancreas.

### Grading & Filtering

Pass (1) **quality** and (2) **correctness** check?

Pass ✅ Fail ❌ → Refining Failed Generation

### Grading & Filtering

Pass (1) **quality** and (2) **correctness** check?

Pass ✅ Fail ❌ → **Regeneration**

## Grading & Filtering

### Step 1: **Quality Check**

**Question:** What is the primary cause of Type 1 Diabetes?
**Answer 1:** A viral infection that specifically targets the pancreas. (**Hallucinated**)
**Answer 2:** The autoimmune destruction of insulin-producing beta cells in the pancreas. (**Ground Truth**)
**Which one is correct?**

| All models got fooled | Two models got fooled | One model got fooled | No model got fooled |
|---|---|---|---|
| ✅ Pass as **HARD** | ✅ Pass as **MEDIUM** | ✅ Pass as **EASY** | ❌ Fail |
| ❌ ❌ ❌ | ● ❌ ❌ | ● ● ❌ | ● ● ● |

### Step 2: **Correctness Check**

If (**Ground Truth Answer** entails **Hallucinated Answer)** AND (**Hallucinated Answer** entails **Ground Truth Answer)**
They have same meaning. -> **Fail** ❌

## Refining Failed Generation

**Failed Hallucinated Answer** → TextGrad → **Improved Answer**

## Fallback After 4 Regeneration Attempts

If the LLM cannot produce a valid hallucinated answer after 5 tries, it will:

- Choose the response that's most similar to the ground truth answer.
- Label this answer as **EASY** ✅