Name _____

Feedback Quality Test - What Helps AI the Most? Answer Key

1. **Wrong."**
   - **Label:** *Somewhat Helpful*
   - **Why?** This tells the AI it made a mistake, but it doesn't say what the correct answer is. The AI knows it was wrong but doesn't know how to fix it.
2. **"That's not correct, it's actually a dog."**
   - **Label:** *Very Helpful*
   - **Why?** This gives a clear correction and the correct label. The AI now knows the right answer and can update.
3. **"Good job!"**
   - **Label:** *Harmful*
   - **Why?** The AI gave the wrong answer, but the human praised it. This reinforces the mistake and makes the AI think "cat" is correct.
4. **"That's wrong. This is a golden retriever, which is a type of dog."**
   - **Label:** *Very Helpful* (best one)
   - **Why?** This gives both the correction and extra detail (dog + golden retriever). The AI learns the general category and the specific type. This is the strongest feedback.
5. **(No response at all.)**
   - **Label:** *Not Helpful*
   - **Why?** The AI gets no signal that it was wrong, so it will keep repeating the same mistake.