# Evaluation

### Goal

The goal of this evaluation is to assess whether the SmartCampus system provides **trustworthy, evidence-backed answers** for campus-related questions. The evaluation focuses on **retrieval quality, grounding, refusal behavior, and risk**, rather than fluent text generation.

### Evaluation Criteria

#### 1. Retrieval Quality

We evaluate whether the system retrieves the correct documents needed to answer a question (e.g., shuttle schedules, campus maps, or policy PDFs).

**Metrics used:**

- Precision@5 (P@5)
- Recall@10 (R@10)

#### 2. Grounding & Faithfulness

We check whether every factual claim in an answer is supported by retrieved evidence.

**Checks include:**

- Correct citation of documents
- No unsupported or fabricated claims
- Evidence matches the answer content

#### 3. Refusal Behavior

If required evidence is missing, the system should **explicitly refuse to answer** instead of guessing.

A correct refusal:

- States that information is missing or unavailable
- Does not fabricate details
- Explains why the answer cannot be given

**4. Human-in-the-Loop Evaluation**

Human reviewers manually inspect system outputs and assign one of the following labels:

- **Yes** – Fully supported by retrieved evidence
- **Partial** – Plausible but missing required evidence
- **Refusal** – Correctly declines due to insufficient evidence

Reviewers also note:

- Missing documents
- Incorrect citations
- Potential real-world risk if deployed

**Baseline Comparison**

System performance is compared against:

- A standard LLM without retrieval
- A RAG system without grounding or refusal rules
- The proposed trust-aware RAG system

**Failure Analysis**

Incorrect or risky outputs are documented to understand system limitations.
Examples include:

- Hallucinated policy rules
- Incorrect shuttle availability
- Missing spatial evidence from campus maps

Each failure is analyzed and paired with a proposed system-level fix.

**Summary**

This evaluation framework emphasizes **trust, transparency, and decision reliability**. By combining automated metrics with human review and explicit refusal behavior, the system is evaluated as a **decision-support tool**, not just a chatbot.