**Model Documentation and Experimental Report**

**1. Model Summary and Description**

The model evaluated is **Mistral-7B-Instruct-v0.3**, a state-of-the-art instruction-tuned transformer with **7 billion parameters**.

* **Architecture**: Decoder-only transformer, optimized for reasoning and natural language understanding.
* **Strengths**: Can generate structured feedback, perform reasoning in technical tasks, and follow prompts consistently.
* **Limitations**: Struggles with subjective or ambiguous human-style responses and shows inconsistent detection ability.

**2. Experimental Setup**

* **Datasets**: Standardized datasets from five domains – *Engineering, Accounting, IT, Psychology, and Teaching*. Each included labeled samples with *AI / Human / Hybrid* labels.
* **Base Prompt**: Unified prompt applied across datasets for consistency.
* **Evaluation Pipeline**:
  + **Detection AI** → Binary classification (AI vs Human).
  + **Feedback AI** → Human-style ratings (1 = poor, 5 = excellent).
* **Metrics**: Detection accuracy (%) and average feedback rating (1–5 scale).

**3. Evaluation Results**

**3.1 Detection Accuracy**

| **Domain** | **Accuracy** |
| --- | --- |
| Engineering | **0.667** |
| Accounting | 0.500 |
| IT | 0.500 |
| Psychology | 0.167 |
| Teaching | 0.500 |

**3.2 Feedback Ratings (1–5)**

| **Domain** | **Mean Score** |
| --- | --- |
| Engineering | 2.66 |
| Accounting | 2.50 |
| IT | **3.66** |
| Psychology | 2.88 |
| Teaching | 2.83 |

**4. Analysis of Results**

**Strengths**

* **Engineering & IT**: Stronger performance — detection accuracy is highest in *Engineering* (66.7%) and feedback quality is strongest in *IT* (3.66). This suggests reliability in structured, technical contexts.
* **Teaching**: Feedback is moderately acceptable (2.83) with balanced detection (50%).

**Weaknesses**

* **Psychology**: Very low detection (16.7%), indicating poor ability to separate human vs AI in subjective/ambiguous content.
* **Accounting**: Both detection (50%) and feedback (2.5) are weak, showing limited reasoning depth in this domain.
* **Feedback Variability**: Scores rarely exceed 4.0, highlighting limited sophistication in nuanced feedback generation.

**5. Explanation of Outcomes**

* **Prompt Sensitivity**: A fixed base prompt restricted adaptability across different subject domains.
* **Model Capability**: Mistral performs better in domains requiring technical reasoning (Engineering, IT) but underperforms in subjective, context-heavy fields (Psychology, Teaching).
* **Detection Challenge**: Distinguishing hybrid responses (partial AI + human) is especially difficult, leading to low accuracy in some domains.

**6. Model Suitability for Project**

* **Feedback AI**:
  + **Suitable** in *Engineering* and *IT* domains (moderate to strong ratings).
  + **Less suitable** for *Psychology* and *Accounting* due to weaker scores.
* **Detection AI**:
  + **Partially suitable** for *Engineering* (best performing).
  + **Unsuitable** in *Psychology* and inconsistent in *Accounting/Teaching*.

**Decision**:

* **Keep** the model for *feedback generation* in structured domains (Engineering, IT).
* **Reject / restrict** for detection tasks in subjective domains (Psychology, Teaching).
* Recommend supplementing detection with a **specialized classifier** or fine-tuning.

**7. Technical Limitations**

* **Domain Generalization**: Limited cross-domain adaptability.
* **Detection Weakness**: Accuracy below acceptable thresholds for high-stakes use.
* **Prompt Dependence**: Results vary significantly with small prompt adjustments.
* **Feedback Ceiling**: Rarely produces ratings above 4, reflecting lack of nuanced evaluation depth.

**8. Conclusion**

The model is **partially suitable** for project integration:

* **Good fit** for generating structured feedback in *Engineering* and *IT*.
* **Not reliable** for detection in *Psychology* and *Accounting*, where nuance and subjectivity dominate.
* **Recommendation**: Use in a **hybrid pipeline** — Mistral for generative feedback, plus a fine-tuned or domain-specific detector for classification.