**Human Evaluation Analysis – AI Detection Task**

**1. Overall Accuracy**

* **Correct Predictions**: 16 / 30
* **Wrong Predictions**: 14 / 30

This is **barely above random chance**, highlighting significant weaknesses in the model’s ability to align with human judgments.

**2. Domain-Level Accuracy**

| **Domain** | **Submissions** | **Correct** | **Wrong** | **Notes** |
| --- | --- | --- | --- | --- |
| **Psychology** | 6 | 3 | 3 | Mixed results, struggles with Hybrid vs Human distinction. |
| **IT** | 6 | 4 | 2 | Best-performing domain. Clear Human essays recognized well. AI misclassified as Human persists. |
| **Engineering** | 6 | 3 | 3 | Similar to Psychology — balance of correct/wrong, mostly errors on Hybrid. |
| **Teaching** | 6 | 2 | 4 | One of the weakest; frequent over-attribution to Human. |
| **Accounting** | 6 | 2 | 4 | Same as Teaching — major misclassification of AI as Human. |

**IT** is the only domain where the detector showed reasonable reliability.  
**Teaching & Accounting** domains are clear weaknesses, dragging down the overall performance.

**3. Error Pattern Analysis**

**False Negatives (AI → Human/Hybrid)**

* **Most frequent error type** (especially in Accounting & Teaching).
* Detector tends to **“trust” AI-like polished academic writing as Human**.
* E.g., **Accounting Submissions 1 & 2** (AI but classified Human).

**Impact:** This undermines the tool’s primary purpose — to **flag AI writing**.

**Human Misclassified (Human → Hybrid/AI)**

* Less frequent but occurs in **Teaching & Accounting**.
* Detector sometimes **overcompensates**, labelling nuanced human writing as “Hybrid.”
* E.g., **Accounting Submission 4** (Human labelled Hybrid).

**Hybrid Misclassifications**

* Hybrid texts are inconsistently detected:
  + Correct in **IT Sub. 6**, **Accounting Sub. 5**
  + Missed in **Engineering, Teaching, and Accounting** (several cases).
* Detector struggles to distinguish between “AI-polished” vs “Human-authored with support.”

**4. Confidence Analysis**

* **High confidence on wrong predictions** → concerning.
* Many AI texts misclassified as Human were given **“High Confidence”** — indicates **systematic blind spot** rather than uncertainty.
* Medium confidence appeared more on Hybrid misclassifications.

**5. Key Insights**

1. **Bias Toward “Human” Labels**
   * When in doubt, the detector leaned Human → leading to **false negatives**.
   * Especially problematic in **Accounting and Teaching domains**.
2. **Domain Variability**
   * IT performed well (accuracy 66.7%).
   * Accounting & Teaching were weakest (33.3% each).
   * Suggests model generalization issues across subject matter.
3. **Hybrid Detection Weakness**
   * Often confused with Human.
   * Needs clearer criteria for Hybrid recognition.
4. **Systematic Overconfidence**
   * Wrong classifications often marked with “High confidence.”
   * Suggests calibration issues in the prediction output.

**6. Recommendations**

* **Recalibrate detection thresholds**: reduce bias toward “Human” in polished, rubric-aligned texts.
* **Improve Hybrid recognition**: train with more nuanced “AI + Human collaboration” samples.
* **Domain-specific fine-tuning**: Accounting & Teaching need specialized training examples.

**Final Summary**

* The detector achieved **53.3% overall accuracy**, with **IT domain strongest (66.7%)** and **Teaching/Accounting weakest (33.3%)**.
* **False Negatives (AI → Human)** are the **largest issue**, showing the detector often fails its main purpose.
* **Hybrid misclassification** further highlights the need for nuanced detection.
* **Overconfidence on wrong predictions** suggests deeper calibration problems.