**TinyLLama Model Evaluation Report**

**AI Detection Performance:**2/5

Critical Performance Issues:

* Barely above random chance with overall performance showing systematic failures
* Severe bias toward Human classification - consistently fails to identify AI-generated content
* High false negative rate - most problematic for academic integrity applications where detecting AI use is essential
* Overconfident incorrect predictions - shows poor calibration between confidence and accuracy
* Inconsistent domain performance - ranges from barely functional to completely unreliable across different academic fields

Domain-Specific Weaknesses:

* Teaching and Accounting domains show severe classification failures
* Only IT domain demonstrates marginally acceptable performance
* Model struggles particularly with Hybrid content detection

**Feedback Generation Performance:**1/5

Fundamental System Failures:

* Cannot execute basic task requirements - fails to provide structured feedback as specified
* Output format violations - returns rubric content instead of assessment feedback
* Incomplete response generation - frequent text cutoffs and truncation issues
* Generic non-specific commentary - lacks ability to provide actionable, targeted feedback
* Text generation degradation - exhibits repetitive token patterns and coherence breakdown
* Complete inability to assess content against rubric criteria

**Overall Assessment**

The TinyLLama model demonstrates critical limitations that render it unsuitable for educational assessment applications.

Key Concerns:

* AI detection reliability is insufficient for academic integrity purposes
* Feedback generation capability is fundamentally broken
* Model shows systematic biases that undermine its core functionality
* Performance inconsistency across domains makes it unreliable for diverse educational contexts

Final Ratings:

* AI Detection: 2/5 - Poor performance with dangerous false negatives
* Feedback Generation: 1/5 - Complete task failure