|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MODEL | RANKING | DIFFER FROM TEACHER | SPELLING SENSITIVITY | TRANSPARENCY | OVERALL FEEDBACK SUMMARY |
| GPT-4o (+o3-mini) | 9.5/10 | +-1 | Low | High | Recognised overall structure; ignored spelling/grammar issues unless it was grading that specific section; very good reference to rubric with excellent reasoning, even with lower prompt-rubric combinations |
| Claude | 9/10 | +-2 | Low | High | Mostly matched teacher grade; similar to gpt, it only penalised spelling in the ‘spelling’ section of the rubric. Overall very good. |
| PaLM | 6.5/10 | +-2/3 | Medium | Medium | Struggled with lower-level rubric/prompt combinations and was inconsistent in applying rubric criteria. Penalised for spelling more often than GPT or Claude. Reasoning was sometimes vague or unsupported. |
| BERT | 6/10 | +-3/4 | High | Low | Frequently over-penalised spelling and grammar, regardless of rubric focus. Showed limited transparency and weak alignment with teacher grades. Lacked depth in justification and often missed rubric cues. |