### **Overview**

We conducted a human evaluation of our **Narrative Auto-Scoring system** using a rubric that covers five key criteria:

1. Establishing Setting, Point of View, and Characters
2. Narrative Techniques
3. Theme/Message
4. Coherence (Organization)
5. Audience Appeal

Evaluators rated the model’s performance across these categories and flagged specific areas where LLM-generated scores or feedback fell short. This analysis highlights patterns of overrating or underrating and offers concrete examples of systematic model behavior.

### **Score Agreement**

We calculated score agreement between LLM and human raters for each criterion (measured as exact match rate):

| **Criterion** | **Score Agreement** |
| --- | --- |
| Setting, Point of View, and Characters | 0.814 |
| Narrative Techniques | 0.876 |
| Theme/Message | 0.857 |
| Coherence (Organization) | 0.832 |
| Audience Appeal | 0.621 |

Audience Appeal had the lowest alignment with human ratings, suggesting a need for prompt refinement in that area.

### **Rating Distribution**

We also tracked how often the LLM *overrated*, *underrated*, or *matched* human scores:

| **Criterion** | **Underrated** | **Overrated** | **Matched** |
| --- | --- | --- | --- |
| Setting, Point of View, and Characters | 12 | 64 | 322 |
| Narrative Techniques | 21 | 21 | 356 |
| Theme/Message | 14 | 36 | 348 |
| Coherence (Organization) | 31 | 26 | 341 |
| Audience Appeal | 175 | 6 | 217 |

Notably, **Audience Appeal** is significantly *underrated*, consistent with qualitative feedback about overly harsh scoring on grammar or tone.

### **Identified Problems by Category**

#### **1. Setting, Point of View, and Characters**

* **Missed setting details**: LLM scored generously even when no setting was established.  
   *Example*: “The student introduces characters, but not the setting, I would give the student a 2.”
* **Ignored voice shifts**: AI didn’t detect inconsistent narration.  
   *Example*: “Who is telling the story? It switches and feels confusing.”

#### **2. Narrative Techniques**

* **Dialogue and description gaps not flagged**: The model didn’t notice when the story lacked vivid elements.  
   *Examples*:  
   “No dialogue and not much in the way of description.”  
   “Dialogue and vivid descriptions and development are missing but feedback doesn’t highlight this clearly.”

#### **3. Theme/Message**

* **Over-credit for vague or missing themes**: LLM assumed the presence of a theme where there wasn’t one.  
   *Examples*:  
   “The student hasn’t developed any message.”  
   “No central message was made clear in the writing.”
* **No support for refining weak messages**: Feedback didn’t guide the student toward sharpening the theme.  
   *Example*: “The message seems implied but not clearly stated.”

#### **4. Coherence**

* **Weak or missing conclusions not noticed** *Example*: “The conclusion is really lacking.”
* **Lack of transition feedback** *Example*: “Transitions would improve the flow and make the ideas more clear.”

#### **5. Audience Appeal**

* **Failed to comment on engagement or voice** *Example*: “There isn't a lot to draw the reader in.”
* **Consistently too harsh with scoring** *Example*: “Voice is consistent and clear — I’d give a 4.”

### **Prompt Revisions**

To address these issues, we have **modified the LLM scoring prompt** with the following adjustments:

* Added instructions to be **more generous on Audience Appeal** scores for younger grades, especially when writing is structured and clear.
* Reinforced the need to **detect missing dialogue, description, or transitions** in Narrative Techniques and Coherence.
* Refined Theme/Message evaluation guidance to **withhold credit when themes are vague or missing**.
* Included checks for **inconsistent narrator voice or unclear perspective** in Setting and Characters.
* Prompt now calls out lack of conclusions and **asks for more specific feedback language** to help students improve.
* Actual changes in the prompt:

6. \*\*For Audience Appeal, Be a little bit generous for lower grades except mostly on copying directly from the passage (e.g., Grades 3–5). If the writing shows effortful structure, readable tone, and mostly clear communication—even with minor errors—assign a 5. Only lower the score if clarity, structure, or tone meaningfully interfere with audience engagement.\*\*

\*\*Additional Reflection Prompts (to improve quality of LLM evaluation):\*\*

- When evaluating setting and characters, ask: \*Do I know where and when this happens? Do I understand who the characters are and how they feel? If not, mention what could be added to make it clearer.\*

- For narrative techniques, ask: \*Did the student use dialogue or description? Did I see any pacing or dramatic effect? If not, suggest adding some.\*

- For theme/message, ask: \*Is the lesson or big idea actually clear? Don’t assume — look for actual lines in the story that express the theme clearly. If you can’t find any, encourage the student to make the message more direct.\*

- When judging organization, ask: \*Does the story have a beginning, middle, and end that flows logically? If the order is confusing or rushed, note that and suggest adding transitions or expanding the conclusion.\*

- When reviewing audience appeal, ask: \*Does the tone and clarity match a {grade} grade student? Avoid scoring too low for small errors, and avoid scoring too high if there are run-ons or confusing language.\*

\*\*Avoid the following issues:\*\*

- Do not reward unclear or vague themes. If the message isn’t stated or supported, score accordingly and explain how to improve.

- Don’t give high scores for narrative technique if there’s little to no dialogue or description.

- Don’t score coherence high if the ending is missing or the order of events is confusing.

- Don’t mark audience appeal too harshly for informal tone unless it disrupts understanding.