Final Project Report

Team Name: group-project-ani-owen-need-one-more

Github repository:

https://github.com/COMPSCI-383-Spring2025/group-project-ani-owen-need-one-more

---

1. Problem Statement

- **Goal:** We evaluate how different prompt structures and languages (English vs. Japanese) affect the quality of short Japanese stories generated by an LLM for JLPT N4 learners.
- **Inputs and Outputs:** Each input is a structured dataset entry containing a theme, characters, vocabulary list, grammar concept, and target character length. The output is a short Japanese story adhering to these.
- **Connection to Requirements:** Our original idea involved predicting cherry blossom bloom dates using historical data. Feedback indicated it lacked depth and didn't showcase generative capabilities, so we pivoted to a more suitable task – evaluating LLM-generated Japanese stories under controlled prompt conditions – better aligned with course's goals on prompt engineering, LLMs and evaluation.

---

2. Dataset

- **Name and Source:** *Japanese N4 Story Dataset* – custom-built.
- **Structure**: 100 entries with:

- 2 characters (sampled from a Japanese name pool)
- 1 theme (from 10 categories)
- 3 JLPT N4 words + 1 out-of-scope vocab
- 1 out-of-scope grammar constraint
- 1 target length (200–300 characters)

- **Dataset Creation or Changes:**
  - Generated using Python script
  - Balanced by theme, character type, and length
  - Saved as dataset.csv and used for automated prompt generation

---

3. Prompt Methodology

- **Prompt Types:**
  - A1: Japanese prompt w/ characters + theme
  - A2: English prompt w/ same
  - A3: Japanese prompt requiring vocab + grammar (incl. out-of-scope)
  - A4: Japanese prompt emphasizing target length
- **Example (A3):**
  - **Prompt:**

    以下の語彙（N4 + 高度な語彙1語）をすべて使って、日本語の物語を書いてください。語彙: 先生、山、話す、論理

    ・JLPT N4レベル以下の語彙、文法、漢字のみを使用してください。

    ・次の高度な文法も1回だけ使用しても構いません: ～ようだ

    ・300文字以内にしてください。

    ・説明や前置きは書かず、物語本文のみを出力してください。

    - **Output (excerpt):**
        - 先生は山の上で話すことが好きだった。論理を使って自然の大切さを教えた。今日は雨が降りそうだ。学生たちは静かに話を聞いた。
- **Sampling Parameters:**
    - Temperature: 0.7
    - Top-p: 0.95
    - Max tokens: 300
    - Model: gpt-4o-mini (OpenAI API)
- **API Call Description:** Responses were generated dynamically and saved to "all_generated_outputs.csv".

---

4. Evaluation Approach

- **Metrics Used:**
    - Coherence: Logical structure of story
    - Grammar Accuracy: Correct N4 or below grammar use and scope control
    - Vocab Accuracy: Use of required vocab, exclusion of forbidden terms out-of-scope
    - Kanji Accuracy: JLPT N4 or below kanji only
    - Length Accuracy: Match target within 10 characters
- **Evaluation Process:**
    - 60 stories sampled (15 per abstraction)
    - Manually scored by two N3-level evaluators (30 each)
    - Metrics scored 0–4 (length: 0–5), max total: 21 points
    - 12 stories double-rated to check inter-rater consistency

- **Strengths and Weaknesses:**
    - **Strengths:**
        - Human review enabled nuanced scoring
        - Diverse, targeted metrics
    - **Weaknesses:**
        - Labor-intensive
        - Limited sample per abstraction due to time

---

5. Results

- **Average Scores (out of 21):**

Abstraction	**Coherence**	**Grammar**	**Vocab**	**Kanji**	**Length**	**Overall Avg**	**Avg -Length**
A1 (JP)	3.40	3.73	3.21	2.53	2.27	**14.93**	**12.88**
A2 (EN)	3.87	4.00	3.53	3.60	5.00	**20.00**	**15**
A3 (Vocab + Grammar)	3.53	3.53	2.33	2.53	4.33	**16.26**	**11.93**
A4 (Length)	3.60	3.87	3.07	3.07	0.67	**14.31**	**13.38**

- **Discussion:**
    - Unexpectedly, A2 (English-prompted) stories scored highest in every single category on average, indicating that GPT-4o-mini can follow Japanese writing constraints even when the prompt is not in Japanese and may even just have better performance when being prompted in English rather than Japanese. Whether this is a typical pitfall of every other language, languages with different

alphabets, or just Japanese would require further research. This performance could also be attributed to the fact that the average length of story was significantly shorter than any other prompt, leaving less room for error but, nonetheless, it did well.

- A3 performed worst, especially when factoring out the length score. Prompting with requests for the use of out of scope vocab and grammar seemed to overload the model, resulting in significantly higher degree of out of scope usage in general as well as more structural and fluency issues. Still, these outputs tended to be more complex and nuanced – indicating that such constraints provide directional guidance, though this often came at the expense of overall accuracy. While we didn't formally assess story complexity, several A3 outputs demonstrated a richer use of transitions and descriptive phrases compared to simpler outputs from A1 and A2.
- A1 produced predictable but sometimes flat stories, suggesting that character and theme constraints alone don't push the model far enough.
- A4 performed better than expected in linguistic categories, though it still suffered when the model struggled to meet or even come near to any one of the stricter character length targets, most of them even reaching lengths higher than any other prompt we tried despite being the prompt specifically testing length constraints
- Overall, these results highlight that prompt density, not language, plays the biggest role in generation quality. The more tightly packed the prompt is with specific requirements, the more the model's performance drops – especially when multiple layers of linguistic control are added at once.

---

6. Feedback and Communication

- **Feedback Received:**
    - Our first proposal lacked complexity and didn't emphasize generative capabilities. It was too narrow in scope and evaluation criteria.
- **How You Addressed It:**
    - We shifted focus to a more complex, generation-driven project involving multilayered constraints (characters, themes, vocab, grammar, length). This pivot allowed deeper exploration of prompt design and evaluation, aligning with course goals.

---

7. Team Member Contributions

- Ani: Dataset generation, prompt scripting, initial draft
- Owen: Evaluation (30 + 6), analysis, editing
- Carey: Evaluation (30 + 6), rubric refinement, final QA