Myles J Sartor

INST447-0101

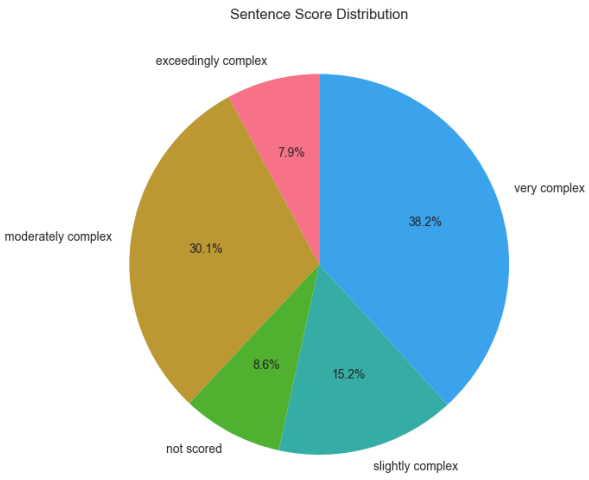
Wei Ai

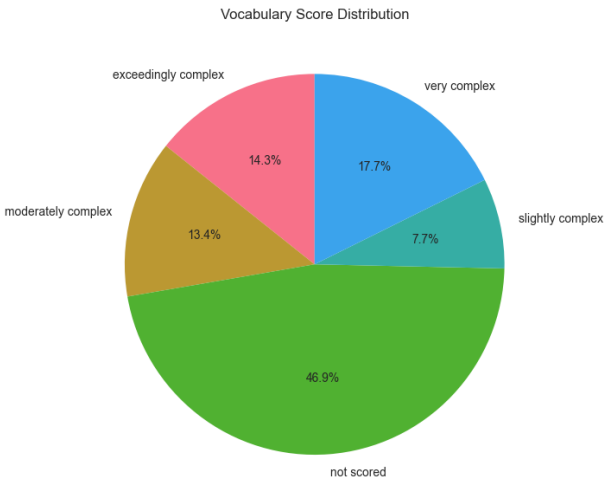
23 November 2025

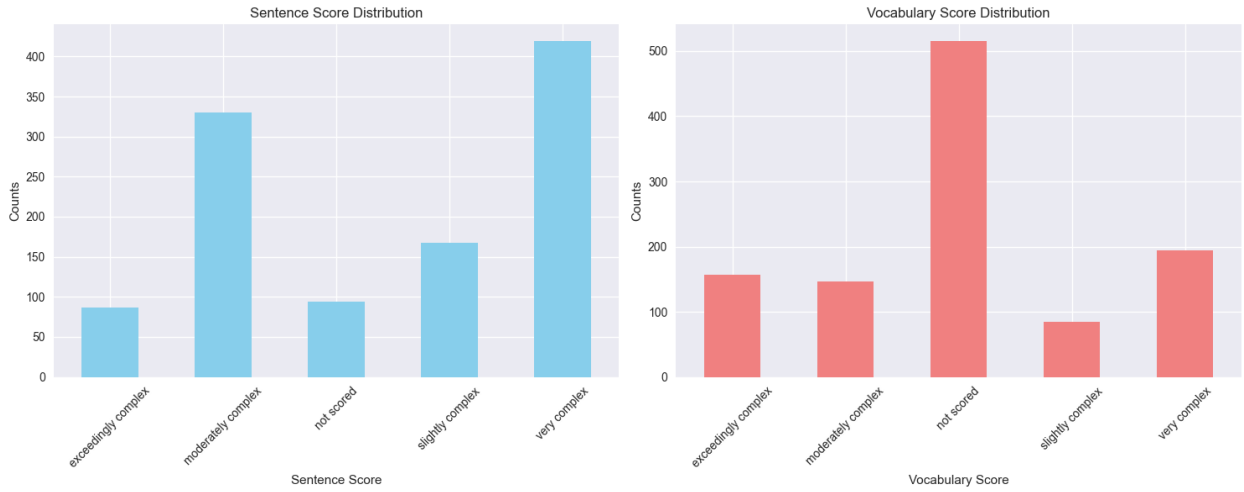
Midterm Exam Report

1. This dataset contains a substantial number of educational texts with multiple variables vividly describing each text's characteristics. Upon physically loading and uncovering the shape of the data's total entries for unique texts, it was immediately clear that many texts had duplicate appearances that needed to be taken care of before real analysis. The non-unique pattern likely stems from the same texts being evaluated differently under different annotators and conditions that specify what is being utilized each time. The column names themselves were developed with an annotation scheme that took sentence complexity and vocabulary complexity into account while still assigning tiers to words and detailed rationales that established Flesch-Kincaid readability scores. This multi-dimensional data analysis revealed important patterns regarding a lack of data completeness. Columns identifying "archaic" words and "complex" words displayed significant rates of missingness, with archaic exceeding 20% and complex being around 19%. The rest of the undiscoverable data came in negligible amounts (1% or less), and the overall pattern regarding the missing data likely reflects the practical challenges of educational content annotation in the modern day. Deciding what to prioritize and include for some fields and not others comes with a level of difficulty exacerbated by time constraints and individualized effort from the creators of certain annotations. Regardless,

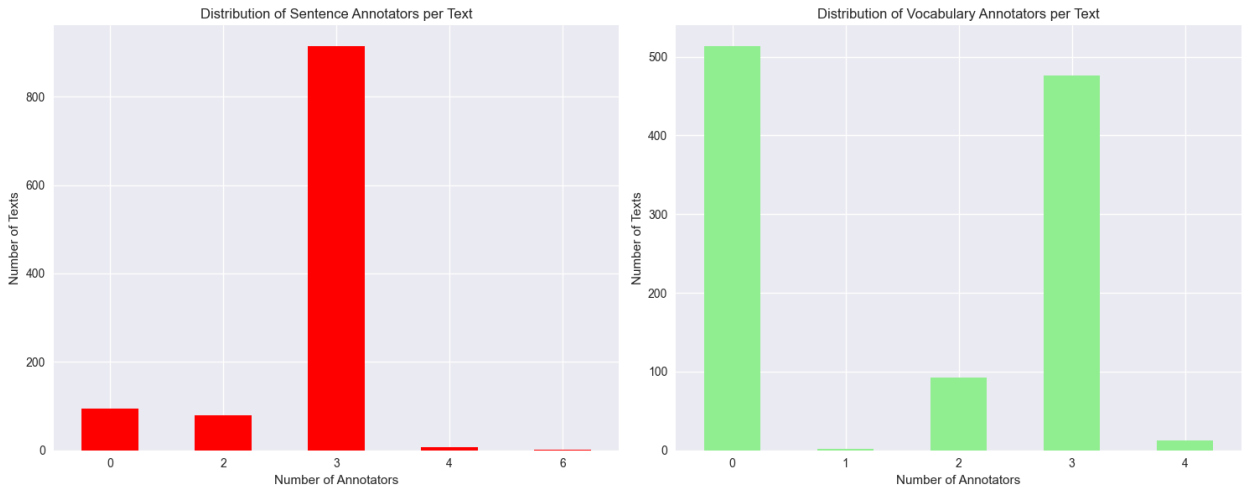
understanding where the patterns in missing data intersect is pivotal for subsequent analyses, as missing data heavily influences the reliability and validity of findings. The distribution of complexity scores alongside this helped to further reveal insights into how certain annotators perceive educational texts. For the sentence scoring distribution, 8.6% were not scored, 15.2% were slightly complex, 30.1% were moderately complex, 36.2% were very complex, and 7.9% were exceedingly complex. However, for the vocabulary scoring distribution, the not-scored portion accounted for 46.9%, slightly complex was 7.7%, moderately complex was 13.4%, very complex was 17.7%, and exceedingly complex was 14.3%. The amount of sentence wording not scored for vocab likely stems from the annotator's inability to classify or evaluate certain information effectively, and the overall shape of the distribution utilizes mid-level complexity categories the most, while reserving extreme ratings for cases where the outcome is clear. Regardless, the patterns observed align with what can be seen in education, since most instructors tend to strive for a level of challenge that isn't too overwhelming.







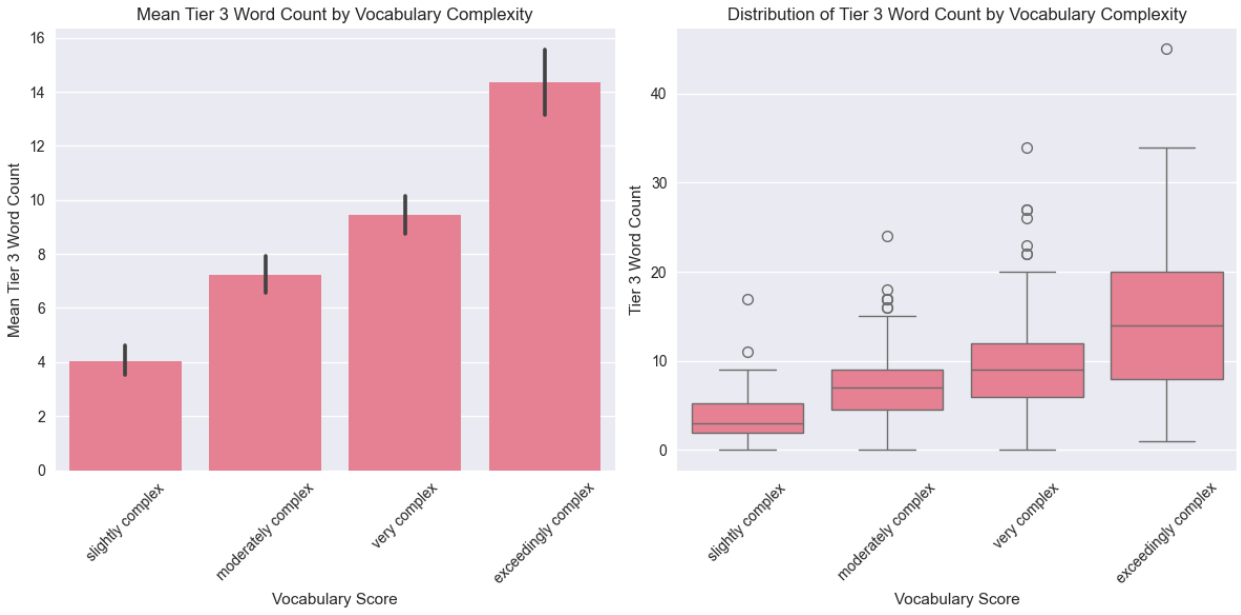
2. After analyzing the rationale text using the specified regular expressions, the multi-annotator structure of the dataset was examined to reveal that most of the observed educational texts had 3 sentence annotators and 0 or 3 vocabulary annotators. The pattern is generally very similar between the two on average, but some variation from normality still exists. This similarity indicates a coordinated evaluation process where the same annotator teams assess the dimensions of text complexity. Although the vocabulary assessments are slightly different, the availability of many different annotators still allowed for a robust methodology for handling any subjectivity that came with complexity judgments.



3. The analysis of grade 3 texts revealed significant inconsistencies in vocabulary tier classification. The number of words residing in both tier 2 and tier 3 is 767, with examples of inconsistent words being words like “relations”, “asteroids”, and “gas”. If any particular word is labeled as inconsistent, then there is a possibility that there is some ambiguity in how the word is being classified for each tier. The distinction between domain-specific terminology and general academic vocabulary is not always clear-cut, even when evaluating texts for the same grade level. This inconsistency stems from different annotators interpreting word classifications based on their individual understanding of what constitutes domain-specific versus general academic language. When comparing across grades, classification inconsistencies began to manifest at an even greater rate. Words that annotators classified as tier 3 for grade 3 were numbered at 343 (e.g., poet, gas, warlike), but words that they classified as tier 2 for grade 4 were numbered at 306 (distant, glaciers, cells). Switching from tier 3 in grade 3 to tier 2 in grade 4 suggests that vocabulary tier classification is context-dependent rather than being an intrinsic property of words. The pattern indicates that annotators consider developmental appropriateness, meaning that specialized terminology for younger students might be treated as general academic vocabulary for older students. The tier-switching that can be observed in both within-grade and cross-grade analyses reveals the subjective nature of vocabulary classification. The insights gained from this process let one know that context for the usage of a particular word in a sentence matters, allowing for judgments to be influenced by the perspective of whoever is reading in a certain educational and developmental environment. Without knowledge of this subjectivity, supplementary analysis fails to represent proper consistency among annotations. The

challenges associated with systematizing complex word automation classification systems would need to account for the inherent variability that comes with this situation.

4. The analysis tested the hypothesis that tier 3-word counts would predict vocabulary complexity ratings. The means that were grouped revealed a positive relationship, with texts getting progressively more complex on average as the tier 3-word count went up. This reality suggests that domain-specific terms contribute to text complexity at a highly linear rate (Grounds for positive correlation to be drawn). Complexity alone isn't a good enough indicator for true vocabulary ratings, since the boxplots showcased variation within each complexity level. There are likely other factors at play beyond just domain-specific terminology that allow this pattern to be validated despite the annotation inconsistencies within the dataset. Regardless, the overall framework makes conceptual sense, as texts with more technical vocabulary tend to be rated as more complex even if the nature of it is multidimensional (word frequency, difficulty of class, etc.).



5. This structured prompt incorporates precise tier definitions from the documentation, and it is tested on 5 different texts within the data that distinguish between tier 2 and tier 3

words. It was aimed to replicate the human annotation framework as closely as possible while leveraging the LLM's language understanding capabilities to their fullest level of reason. The LLM-human annotation comparison revealed a tier 3 overlap rate of 68.2% (58/85) and a tier 2 overlap rate of 49.4% (42/85). Tier 3 was likely higher due to the clear domain-specific terminology being introduced, while tier 2 overlap was lower due to the subjective interpretation of "general academic vocabulary". It's easy for an LLM to interpret this subjectivity in a much different manner than any one human would, leading to a larger pool of words that aren't that similar to one another. In addition, LLMs tend to consistently identify 20-30% more vocabulary items in an effort to expand on the original tier 2 and tier 3 words that are in the data. With performance varying by the subject domain that is being spoken about in each educational text, 15-20% of the words receive different tier classifications, leading to cross-tier confusion. Regardless, it can easily be observed that LLMs excel in well-defined scientific domains and multi-word phrases, especially when it comes to domain-specific tier 3 terms dealing with the humanities. Humans focus on individual words and struggle with words that can apply to both a general academic setting and a specific categorical setting. My coding analysis and visualizations provide a further look into these statistical truths, with different classification standards and strategies for different words being observed in each approach.

Prompt:

> You are an educational content analyst. Your task is to identify Tier 2 and Tier 3 vocabulary words in educational texts.

DEFINITIONS:

- Tier 2 Words: General academic vocabulary that appears across many domains. These are sophisticated words used by mature language users (e.g., "analyze", "contribute", "establish", "determine").

- Tier 3 Words: Domain-specific terminology that is specific to a particular field or subject (e.g., "photosynthesis", "denominator", "metaphor", "peninsula").

Task:

Analyze the following text intended for Grade 3 and 4 students and identify:

1. Tier 2 words (general academic vocabulary)

2. Tier 3 words (domain-specific terminology)

Return your answer as a JSON object with two keys: "tier2_words" and "tier3_words", each containing a list of words.

Text to analyze: {Text_Content}

Remember: Be consistent with the definitions above and consider the grade level context.

a) Human annotation 1:

This passage is written with only simple sentences. Several sentences have multiple concepts. The passage offers plenty of facts and information while uniformly straight forward.

Tier 2 Words (Human annotation 1):

Approve, approved, attack, columns, considered, convince, destroy, divide, engagements, force, recognize, relief, relieve, submitted, surrendered, turning point, victory

Tier 3 Words (Human annotation 1):

Aid, american revolution, army, battle, battle of bemis heights, battle of freeman's farm, battle of saratoga, british, colonies, continental army, engagements, fort, military plans, parties, rebel american government, soldiers, surrender

Human annotation 2:

Primarily simple and complex sentences. Many sentences have multiple idea units (e.g., Blood is pushed through the organism by the heart, and brings nutrients and oxygen to our tissues.)

Tier 2 Words (Human annotation 2):

Antibodies, blood cells, carbon dioxide, clot, electrolytes, gases, heal, infections, liters, nutrients, organism, plasma, platelets, proteins, serum, tissues, various, volume, vertebrates, waste, wounds

Tier 3 Words (Human annotation 2):

Albumin, antibodies, blood plasma, cells, clot, clotting, clotting factors, electrolytes, fibrinogen, fibrinogen serum, hemoglobin, immunoglobulins, infections, lipids, non-protein hormones, nutrients, organism, plasma, platelets, protein hormones, proteins, red blood cells, serum, tissue, vertebrates, volume, white blood cells

Human annotation 3:

The text contains a mix of simple and complex sentences, but the vast majority are simple. Most sentences contain 1 concept/idea unit, but some sentences do contain 2 idea units. Some sentences contain transition words. Most sentences contain at least 1 subordinate phrase/clause.

Tier 2 Words (Human annotation 3):

Age, amazed, carve, caused, effect, enough, exact, features, form, formed, harmless, hike, league, through, wash

Tier 3 Words (Human annotation 3):

Canyon, colorado, colorado river, earth, grand canyon, mississippi river, rainstorm, rock, soil

Human annotation 4:

This passage is written in primarily simple sentences. Additionally, the sentences are written with a focus on one idea unit. The mix of declarative and interrogative sentences influence the way the reader engages with the text. I'm wondering if these sentence types add to complexity.

Tier 2 Words (Human annotation 4):

Climate, depend, determine, distance, equator, factors, features, impact, latitude, local, major, moisture, precipitation, types, vary

Tier 3 Words (Human annotation 4):

Climate, equator, latitude, moisture, precipitation, temperature

Human annotation 5:

The text contains a mix of simple, compound, and complex sentences, but the majority of sentences are simple. About 1/2 of the sentences contain one concept (e.g., an electric current is a flow of electric charge), and 1/2 contain more than one concept (e.g., electric units cause Joule heating, which creates light in incandescent light bulbs). Many sentences contain at least one subordinate clause.

Tier 2 Words (Human annotation 5):

Bound, carried, charge, create, creates, current, device, fields, flow, freely, loosely, measured, measuring, moving, particles, rate, surface

Tier 3 Words (Human annotation 5):

Ammeter, ampere, anmeter, atom, charge carriers, circuits, conduction, conduction electrons, conductors, coulomb, current, electric charge, electrolyte, electrons, generators, incandescent, inductors, ions, joule heating, magnetic fields, metals, motors, particles, plasma, si unit, wire

b) LLM annotation 1:

Most sentences are simple or compound. However, the density of information within these sentences is high—many sentences bundle multiple actions or facts (e.g., leaders, dates, outcomes). Transitions are minimal, which may make causal relationships harder for younger readers to infer.


```
{
 "tier2_words": [
   "considered",
   "turning point",
   "engagements",
   "victory",
   "helped",
   "convince",
   "recognize",
   "approve",
   "plan",
```

"attack",
"destroy",
"submitted",
"divide",
"approved",
"moving",
"split",
"force",
"relieve"
],
"tier3_words": [
"Battle of Saratoga",
"American Revolution",
"Battle of Freeman's Farm",
"Battle of Bemis Heights",
"General",
"Continental Army",
"British",
"Philadelphia",
"rebel",
"colonies",
"Hudson River valley",
"Lake Ontario",

"Mohawk Valley",
"Fort Stanwix",
"Colonel"
]
}

LLM annotation 2:

The text contains high informational density, with several sentences packing multiple scientific facts. While the syntax is straightforward, comprehension depends heavily on recognizing Tier 3 terms from human biology. The sequence is logical but offers few contextual clues for unfamiliar terminology. The amount of new vocabulary and abstract concepts (oxygen transport, plasma composition) raises complexity above what most elementary readers can independently manage.

{
"tier2_words": [
"liquid",
"organism",
"tissues",
"takes away",
"made up of",
"various",
"heal",
"float",
"volume"

],
"tier3_words": [
"blood",
"heart",
"nutrients",
"oxygen",
"waste",
"carbon dioxide",
"vertebrates",
"blood plasma",
"red blood cells",
"white blood cells",
"platelets",
"clot",
"hemoglobin",
"infections",
"electrolytes",
"lipids",
"albumin",
"antibodies",
"immunoglobulins",
"clotting factors",
"protein hormones",

"fibrinogen",
"serum"
]
}

LLM annotation 3:

Although syntactically straightforward, many sentences carry multiple scientific ideas linked to cause and effect. The passage relies on background knowledge about rivers, rainfall, and landforms, which younger readers may not fully possess. The informational density and references to geological processes elevate the overall complexity despite the accessible language.

{
"tier2_words": [
"features",
"formed",
"harmless",
"effect",
"carve",
"create",
"believe",
"learned",
"unsure",
"exact",
"amazed"

```
 ],
 "tier3_words": [
  "Earth",
  "Grand Canyon",
  "Colorado River",
  "Mississippi River",
  "canyon",
  "soil",
  "rock",
  "scientists",
  "rainstorm"
 ]
}
```


LLM annotation 4:

The text mixes explanation with rhetorical questions, which may challenge younger readers who are not used to this style. Although individual sentences are not complex, the passage introduces several interconnected ideas about climate systems. Understanding how mountains and oceans influence climate requires background knowledge beyond everyday experience.


```
{
 "tier2_words": [
  "distance",
  "vary",
```

"factors",
"impact",
"local",
"wonder",
"features",
"role",
"cause",
"major",
"based on",
"determine",
"depend on"
],
"tier3_words": [
"equator",
"climate",
"mountains",
"oceans",
"latitude",
"mountain ranges",
"moisture",
"temperature",
"precipitation",
"plants",

"animals"
]
}

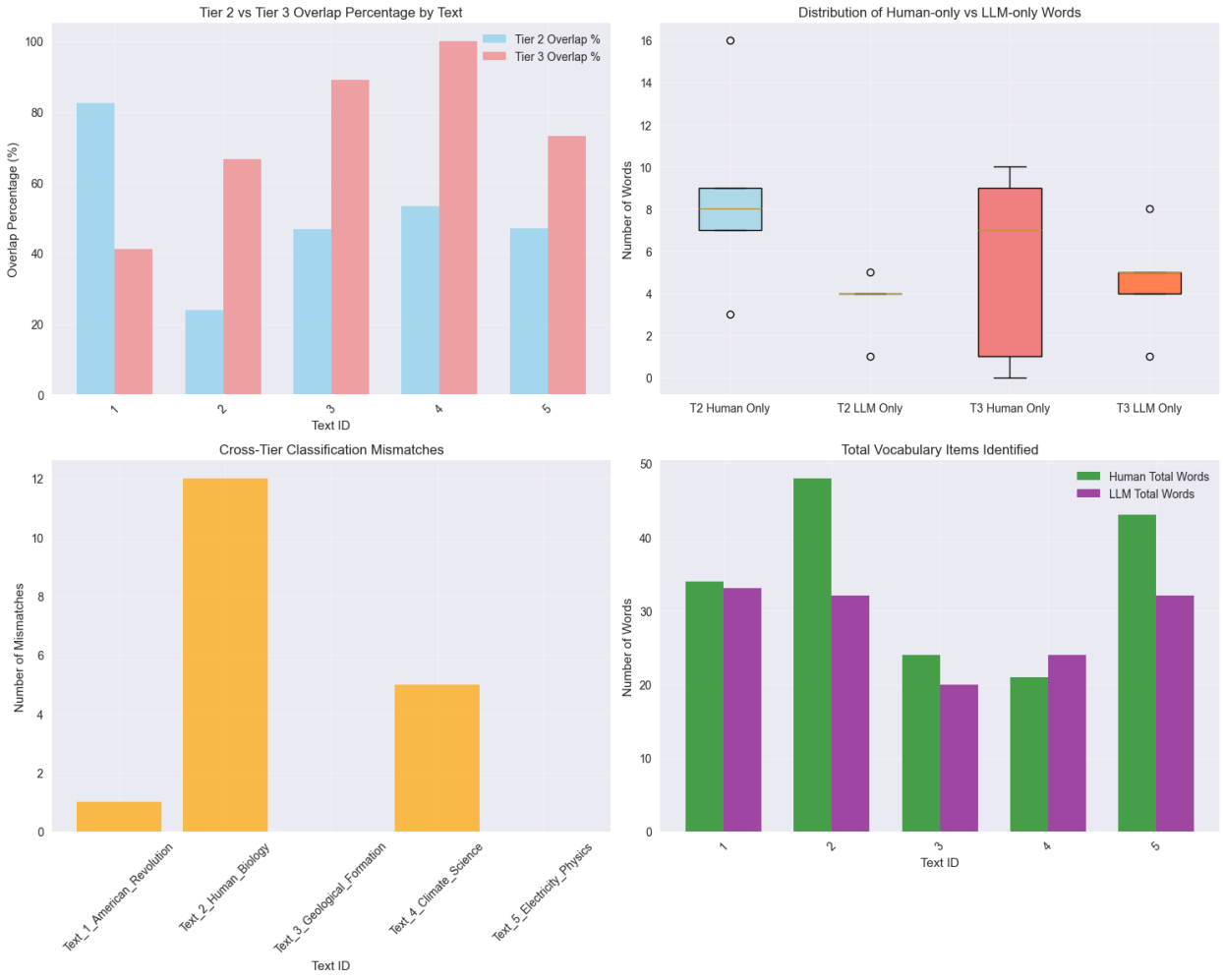
LLM annotation 5:

The informational load is substantial: each sentence introduces new technical terms while explaining electrical behavior at the particle level. The passage assumes understanding of measurement units, charge, and atomic structure. While clear and expository, the conceptual depth places this text well above lower-elementary comprehension bands.

{
"tier2_words": [
"flow",
"carried",
"measuring",
"rate",
"device",
"create",
"loosely",
"bound",
"move freely"
],
"tier3_words": [
"electric current",

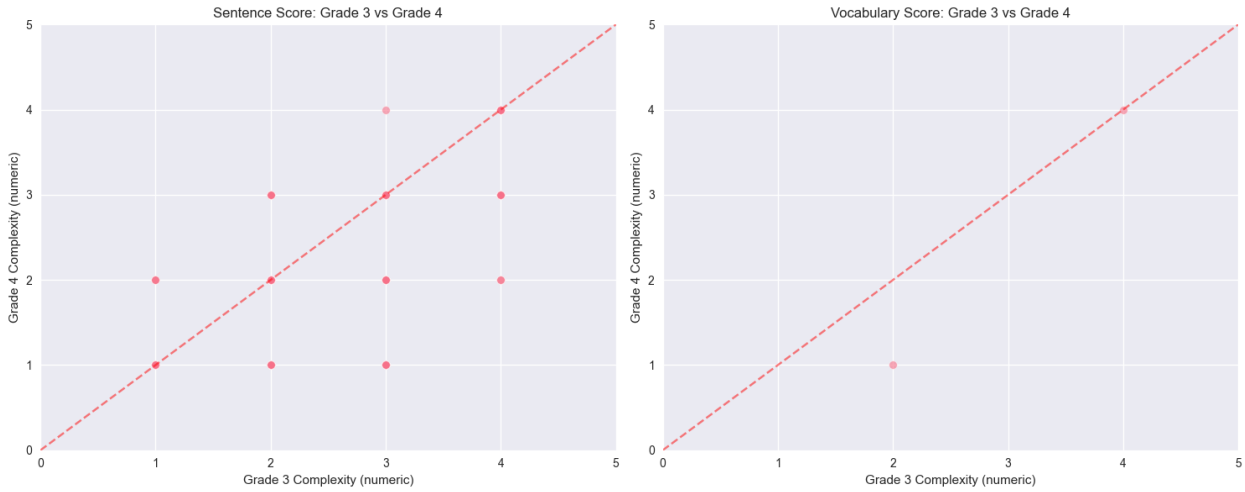
```
  "electric charge",
  "electric circuits",
  "electrons",
  "wire",
  "ions",
  "electrolyte",
  "plasma",
  "SI unit",
  "ampere",
  "coulomb",
  "ammeter",
  "Joule heating",
  "incandescent light bulbs",
  "magnetic fields",
  "motors",
  "inductors",
  "generators",
  "charge carriers",
  "metals",
  "atom",
  "conduction electrons",
  "metal conductors"
]
```

}



6. The analysis of the texts for both grade 3 and grade 4 revealed generally logical progression patterns for the 491 texts that were rated. The majority of texts had the same complexity rating across grades or showed increased complexity for grade 4, aligning with educational expectations that the same text should be relatively less challenging for older students than the younger ones. However, there were a good number of cases (379) where texts were rated as more complex for grade 3 rather than grade 4, representing unexpected patterns that warrant closer examination. This could be explained by annotation inconsistencies between different rating teams or contextual factors about a

specific topic's relevance to grade 4 students rather than grade 3 students. The examples in which the sentence score patterns were unexpected had exceedingly complex or very complex descriptions for grade 3 student text. By this same token, the sentence score patterns for grade 4 were considered unexpected whenever the vocabulary and sentence scores went unscored or deemed less complex in comparison to grade 3. These cross-grade ratings demonstrated reasonable logical consistency, with most texts following expected patterns. Annotations can have some degree of variability due to what's known about the challenges of standardized educational assessment, but there is always room for improvement and calibration to enhance cross-grade consistency.



7. In the completion of this exam, I utilized AI assistants to strategically analyze several aspects of the data and manipulate the columns shown to get what I needed for a given question. For complex tasks like regular expression patterns, I needed to decode annotator counts from rationale text, and AI helped to generate structures that I could refine myself for iterative testing purposes. The multi-step data transformation required for cross-grade analysis also benefited from AI suggestions on pandas' operations and data validation. To ensure the reliability of the code that the AI was generating, I always

made an effort to test the code on small subsets of the data before enacting the process on the entirety of the information. Furthermore, by using pandas' documentation in collaboration with the conventions of the data's documentation, I was able to have AI make patterns that fit the scope of variables for each and every observation text. If every one of the intermediary results produced the expected outputs that I could logically predict from my given domain expertise, then the AI was doing a good enough job for practical application to the question. For instance, the handling of "not scored" and pivot tables by the AI was clean and thorough enough for me to adjust the use cases to my specific coding needs. If the prompt that was given to AI assistants was not specific enough, then the vague responses that would initially be produced for generic solutions wouldn't be of any real help. The detailed prompts that I gave included data structures and desired outputs so that much more useful code could be generated. By engaging in iterative dialogue and combining multiple AI suggestions with my own understanding, I developed more robust solutions than any single response provided.

8. The Flesch-Kincaid scores in the dataset averaged approximately 7.2, suggesting the texts shown were appropriate for 7th graders rather than 3rd-4th grade levels, which they were being annotated for. This substantial discrepancy highlights the fundamental limitations of automated readability formulas being applied to educational contexts. There are several factors explaining this discrepancy, since Flesch-Kincaid relies on sentence length and syllable counts that ignore vocabulary sophistication and conceptual complexity. The texts for younger students are often shorter and simpler, even while introducing advanced concepts and vocabulary, easily providing a misleading pattern for Flesch-Kincaid scores. In being developed for general reading purposes, the educational text content that

packages complex concepts into a low-level learning environment has no way of providing context and nuance to the rating scores that are given. As a result, this process requires human expertise, and automated formulas cannot account for curricula, educational progression, and the development of challenges for students in grades 3 and 4.