1 KAI Evaluation Framework - Demo

KAI is scaling from a single RCT to three concurrent pilot contexts: US public schools, Odisha India, and a parent-facing home version. Manual rubric scoring (the methodology that gave the original RCT its credibility) does not scale across these contexts.

This notebook prototypes one piece of that scaling problem: an automated rubric-based evaluator grounded in the Lemons/Balasubramanian research group's science-of-reading framework, designed specifically for student populations with intellectual and developmental disabilities (IDD).

[1]:
```
# running locally
from dotenv import load_dotenv
load_dotenv()
```

[1]:
```
True
```

1.1 Setup

If you're running this in Google Colab: 1. Run the cell below to install dependencies 2. Paste your Anthropic API key when prompted

If you're running locally, set `ANTHROPIC_API_KEY` in your environment or in a `.env` file.

[2]:
```
# installing dependencies and setting up the api key
import sys
import os

# installing packages if running in colab
if 'google.colab' in sys.modules:
    !pip install -q anthropic python-dotenv
    !git clone -q https://github.com/YOUR_USERNAME/kai-eval-framework.git 2>/dev/
 ↪null || true
    sys.path.insert(0, 'kai-eval-framework')
else:
    # running locally - adding parent directory to path
    sys.path.insert(0, os.path.abspath('..'))

# getting the api key
if not os.environ.get('ANTHROPIC_API_KEY'):
    from getpass import getpass
    os.environ['ANTHROPIC_API_KEY'] = getpass('Enter your Anthropic API key: ')

print('setup complete.')
```

```
setup complete.
```

1.2 The Rubric

The rubric below is grounded in the science-of-reading framework and the published work of the Lemons and Balasubramanian research group on multicomponent reading intervention for students

with IDD. I leaned specifically on Heidlage et al. (2024) (the parent-implemented intervention paper) because its framing of how children with IDD demonstrate literacy gains directly informs how the rubric should (and shouldn't) score responses.

The most important design decision is the **response validity** dimension. Students with IDD often communicate in ways that look non-standard on the surface like incomplete sentences, unconventional grammar, phonetic spelling, word order that doesn't match formal English. A naive evaluator would score these responses as low quality.

But that conflates *how* a student communicates with *whether* they've understood the passage, and those are different things. The rubric explicitly instructs the judge to separate the two and to honor the presumed competence principle that anchors the team's broader research.

[7]:
```
# displaying the rubric
from src.rubric import RUBRIC
from IPython.display import display, Markdown

# mapping names to formal display titles
DISPLAY_NAMES = {
    "literal_comprehension": "1. Literal Comprehension",
    "main_idea_identification": "2. Main Idea Identification",
    "inferential_reasoning": "3. Inferential Reasoning",
    "response_validity": "4. Response Validity",
}

for dim in RUBRIC.dimensions:
    display_name = DISPLAY_NAMES.get(dim.name, dim.name.replace("_", " ").
 ↪title())
    md = f"### {display_name}\n\n"
    md += f"{dim.description}\n\n"
    md += "| Score | Description |\n|-------|-------------|\n"
    for score, desc in dim.score_anchors.items():
        md += f"| {score} | {desc} |\n"
    if dim.critical_notes:
        md += f"\n> **Critical:** {dim.critical_notes}\n"
    md += "\n---\n"
    display(Markdown(md))
```

1.2.1 1. Literal Comprehension

Did the student accurately identify information directly stated in the passage? Measures whether the student can locate and reproduce factual content from the text.

Score	Description
0	No accurate information from the passage is present in the response.
1	One piece of accurate information is present, but it may be vague or incomplete.

Score	Description
2	Multiple accurate details from the passage are present and mostly correct.
3	The response accurately captures key factual content directly stated in the passage.

1.2.2 2. Main Idea Identification

Did the student capture the central point of the passage, not just surface details? This is the primary outcome measure in the KAI RCT.

Score	Description
0	The response does not address the main idea of the passage.
1	The response touches on a related topic but misses the central point.
2	The response captures part of the main idea but is incomplete or includes irrelevant details.
3	The response clearly and accurately identifies the central point of the passage.

1.2.3 3. Inferential Reasoning

Did the student draw appropriate inferences beyond what is directly stated? Tests comprehension beyond sight-word recognition, which is historically the only skill IDD students were taught.

Score	Description
0	No inference is present — the response only restates or is off-topic.
1	A weak or partially correct inference is present.
2	A reasonable inference is present but may lack depth or precision.
3	The response demonstrates a clear, well-supported inference that goes beyond the text.

1.2.4 4. Response Validity

Does the response demonstrate genuine comprehension, accounting for non-standard communication patterns? Non-standard grammar, spelling, incomplete sentences, or unconventional phrasing do NOT lower this score. Only comprehension content matters.

Score	Description
0	The response does not demonstrate any comprehension of the passage.
1	Minimal comprehension is present, but it is unclear or ambiguous.

Score	Description
2	The response demonstrates partial comprehension despite any surface-level irregularities.
3	The response demonstrates clear comprehension. Surface-level irregularities (grammar, spelling, sentence structure) are irrelevant to this score.

> **Critical:** This dimension encodes the 'presumed competence' principle. Non-standard grammar, spelling, incomplete sentences, phonetic spelling, unconventional word order, or any other surface-level irregularity must NOT reduce this score. The ONLY thing that matters is whether the student has demonstrated comprehension of the passage content. An evaluator that penalizes non-standard communication has failed its most important test.

1.3 The Evaluator

The evaluator runs two scoring passes at different temperatures (0.1 and 0.4) and compares them. If any dimension's score differs by more than 1 point between passes, the response is flagged as uncertain and recommended for human review. Each pass requires the judge to cite specific phrases from the student's response as evidence - a grounding constraint that reduces hallucinated scoring. The output is structured JSON with per-dimension scores, reasoning, evidence phrases, and a confidence level.

1.4 Running the Evaluator on Synthetic Examples

The following examples are synthetic but designed to stress-test the evaluator on cases that matter for IDD populations. They range from strong comprehension to off-topic responses, with a critical test case that probes whether the evaluator honors the presumed competence principle.

[9]:
```
# loading the examples
from src.examples import EXAMPLES
from IPython.display import display, Markdown

# mapping internal ids to clean display titles
EXAMPLE_TITLES = {
    "photo_strong": "Photosynthesis: Strong Comprehension",
    "photo_surface": "Photosynthesis: Surface-Level Response",
    "photo_offtopic": "Photosynthesis: Off-Topic Response",
    "photo_critical": "Photosynthesis: Critical Test Case",
    "stamp_inferential": "Stamp Act: Strong Inferential Reasoning",
    "stamp_literal": "Stamp Act: Literal Regurgitation",
}

for ex in EXAMPLES:
    title = EXAMPLE_TITLES.get(ex['id'], ex['id'])
    # cleaning em-dashes from descriptions
    description = ex['description'].replace('-', ':').replace('-', ':')
```

```
md = f"### {title}\n\n"
md += f"**Description:** {description}\n\n"
md += f"**Question:** {ex['question']}\n\n"
md += f"**Student Response:** \"{ex['student_response']}\"\n\n---\n"
display(Markdown(md))
```

1.4.1 Photosynthesis: Strong Comprehension

Description: High-quality response : student captures the main idea and shows inferential reasoning about the plant-sun relationship.

Question: What is the main idea of this passage?

Student Response: "The main idea is that plants make their own food using sunlight through photosynthesis. They need sunlight, water, and carbon dioxide, and this happens in their leaves where chlorophyll is. Without this process, plants and most other living things couldn't survive because they all depend on it."

1.4.2 Photosynthesis: Surface-Level Response

Description: Partial response : student identified a surface feature but did not capture the central point.

Question: What is the main idea of this passage?

Student Response: "Plants are green and they need water to grow."

1.4.3 Photosynthesis: Off-Topic Response

Description: Off-topic response : student did not engage with the passage content.

Question: What is the main idea of this passage?

Student Response: "I like flowers in my garden. My mom has roses and they are red."

1.4.4 Photosynthesis: Critical Test Case

Description: CRITICAL TEST CASE : non-standard grammar, incomplete sentences, but fundamentally correct comprehension. The evaluator must score this HIGH on literal_comprehension, main_idea_identification, and response_validity. Any evaluator that penalizes this response for grammar has failed the presumed competence test.

Question: What is the main idea of this passage?

Student Response: "plant make food from sun they need water too leaves help them grow"

1.4.5 Stamp Act: Strong Inferential Reasoning

Description: Strong inferential reasoning : student connects cause and effect beyond what is directly stated.

Question: Why did the colonists protest the Stamp Act?

Student Response: "The colonists protested because they thought it was unfair to be taxed when they had no say in the government making the rules. They didn't have anyone representing them in Parliament, so they felt like the British government was taking their money without giving them a voice. That's why they boycotted and formed groups to fight back."

1.4.6 Stamp Act: Literal Regurgitation

Description: Literal regurgitation : student accurately pulled information from the passage but did not demonstrate inferential reasoning.

Question: Why did the colonists protest the Stamp Act?

Student Response: "The Stamp Act was passed in 1765 and it made colonists pay tax on newspapers and legal documents. It was repealed in 1766."

[10]:
```
# scoring all examples
from src.evaluator import score_response
from IPython.display import display, Markdown

# clean display names for dimensions
DIMENSION_DISPLAY = {
    "literal_comprehension": "Literal Comprehension",
    "main_idea_identification": "Main Idea Identification",
    "inferential_reasoning": "Inferential Reasoning",
    "response_validity": "Response Validity",
}

results = {}

for ex in EXAMPLES:
    title = EXAMPLE_TITLES.get(ex['id'], ex['id'])
    print(f"scoring: {title}...")

    result = score_response(
        passage=ex['passage'],
        question=ex['question'],
        student_response=ex['student_response'],
    )
    results[ex['id']] = result

    # displaying results
    md = f"### Results: {title}\n\n"
    md += f"**Student said:** \"{ex['student_response']}\"\n\n"

    if result.flag:
        md += f" **Flag:** {result.flag}\n\n"

    md += "| Dimension | Score | Reasoning␣
↪|\n|-----------|-------|-----------|\n"
```

```
    for dim_name in result.scores:
        display_name = DIMENSION_DISPLAY.get(dim_name, dim_name)
        score = result.scores.get(dim_name, 'N/A')
        reasoning = result.reasoning.get(dim_name, '')
        # truncating reasoning for table readability
        if len(reasoning) > 120:
            reasoning = reasoning[:120] + '...'
        md += f"| {display_name} | {score} | {reasoning} |\n"

    md += f"\n**Agreement between passes:** {result.agreement}\n\n---\n"
    display(Markdown(md))

print("all examples scored.")
```

```
scoring: Photosynthesis: Strong Comprehension...
```

1.4.7 Results: Photosynthesis: Strong Comprehension

Student said: "The main idea is that plants make their own food using sunlight through photosynthesis. They need sunlight, water, and carbon dioxide, and this happens in their leaves where chlorophyll is. Without this process, plants and most other living things couldn't survive because they all depend on it."

Dimension	Score	Reasoning
Literal Comprehension	3	The student accurately identifies multiple factual details directly stated in the passage: the process name (photosynthe...
Main Idea Identification	3	The student clearly and accurately identifies the central point of the passage: that plants make their own food through ...
Inferential Reasoning	3	The student demonstrates inferential reasoning by connecting the survival of plants to the survival of 'most other livin...
Response Validity	3	The response demonstrates complete and clear comprehension of the passage. The student has synthesized information from ...

Agreement between passes: True

```
scoring: Photosynthesis: Surface-Level Response...
```

1.4.8 Results: Photosynthesis: Surface-Level Response

Student said: "Plants are green and they need water to grow."

Dimension	Score	Reasoning
Literal Comprehension	1	The student identifies two factual details that are present or implied in the passage: that plants are green (related to...
Main Idea Identification	1	The passage's main idea is that plants make their own food through photosynthesis, a process essential for plant surviva...
Inferential Reasoning	0	The response does not demonstrate any inferential reasoning beyond what is directly stated or visually obvious. The stat...
Response Validity	1	The response demonstrates minimal comprehension of the passage content. While the student has identified two accurate de...

Agreement between passes: True

`scoring: Photosynthesis: Off-Topic Response...`

1.4.9 Results: Photosynthesis: Off-Topic Response

Student said: "I like flowers in my garden. My mom has roses and they are red."

Dimension	Score	Reasoning
Literal Comprehension	0	The response does not contain any information from the passage. The passage discusses photosynthesis, sunlight, water, c...
Main Idea Identification	0	The main idea of the passage is that plants make their own food through photosynthesis, and this process is essential fo...
Inferential Reasoning	0	The response does not demonstrate any inference related to the passage content. While the student mentions plants (flowe...
Response Validity	0	The response does not demonstrate comprehension of the passage. While the response is clearly written and communicates t...

Agreement between passes: True

`scoring: Photosynthesis: Critical Test Case...`

1.4.10 Results: Photosynthesis: Critical Test Case

Student said: "plant make food from sun they need water too leaves help them grow"

Dimension	Score	Reasoning
Literal Comprehension	3	The student accurately identifies multiple factual details directly stated in the passage: that plants make food ('plant...
Main Idea Identification	3	The student clearly captures the central point of the passage: plants make their own food through a process involving su...
Inferential Reasoning	2	The student demonstrates inferential reasoning by connecting the process to growth ('leaves help them grow'), which requ...
Response Validity	3	The response demonstrates clear and genuine comprehension of the passage content. The student has identified the core co...

Agreement between passes: True

`scoring: Stamp Act: Strong Inferential Reasoning...`

1.4.11 Results: Stamp Act: Strong Inferential Reasoning

Student said: "The colonists protested because they thought it was unfair to be taxed when they had no say in the government making the rules. They didn't have anyone representing them in Parliament, so they felt like the British government was taking their money without giving them a voice. That's why they boycotted and formed groups to fight back."

Dimension	Score	Reasoning
Literal Comprehension	3	The student accurately identifies multiple factual details directly stated in the passage: the colonists had no represen...
Main Idea Identification	3	The student clearly captures the central point of the passage: the colonists protested because they were being taxed wit...
Inferential Reasoning	3	The student demonstrates inferential reasoning by connecting the concept of 'no representation' to the broader principle...

Dimension	Score	Reasoning
Response Validity	3	The response demonstrates clear and comprehensive comprehension of the passage. The student accurately answers the quest...

Agreement between passes: True

`scoring: Stamp Act: Literal Regurgitation...`

1.4.12 Results: Stamp Act: Literal Regurgitation

Student said: "The Stamp Act was passed in 1765 and it made colonists pay tax on newspapers and legal documents. It was repealed in 1766."

Dimension	Score	Reasoning
Literal Comprehension	2	The student accurately recalls multiple factual details directly stated in the passage: the year the Stamp Act was passe...
Main Idea Identification	0	The question specifically asks WHY the colonists protested the Stamp Act. The central point of the passage is that colon...
Inferential Reasoning	0	The response does not demonstrate any inferential reasoning. It only restates factual information directly from the pass...
Response Validity	1	The response demonstrates minimal comprehension. While the student accurately recalls some factual details about the Sta...

Agreement between passes: True

`all examples scored.`

1.5 The Critical Test Case

This is the most important example in the set. Example `photo_critical` is a student response with non-standard grammar, incomplete sentences, and no punctuation:

> "plant make food from sun they need water too leaves help them grow"

This response demonstrates **correct comprehension** of photosynthesis. The student identifies that plants make food, that they use sunlight and water, and that leaves are involved. Every key concept from the passage is present.

If the evaluator scores this response low on `literal_comprehension`, `main_idea_identification`, or `response_validity`, **it has failed its most important test.** An evaluator that penalizes non-standard grammar is not measuring comprehension — it's measuring writing ability, and those are different things. The presumed competence principle requires us to separate the two.

Let's look at how the evaluator handled it:

[11]:
```
# examining the critical test case
from IPython.display import display, Markdown

critical = results.get('photo_critical')

if critical is None:
    print("critical example not scored - check for errors above.")
else:
    md = "### Critical Test Case Results\n\n"
    md += "**Student response:** \"plant make food from sun they need water too␣
 ↪leaves help them grow\"\n\n"

    # checking if the evaluator passed the test
    passed = True
    key_dims = ['literal_comprehension', 'main_idea_identification',␣
 ↪'response_validity']
    for dim in key_dims:
        score = critical.scores.get(dim)
        if score is not None and score < 2:
            passed = False

    if passed:
        md += "**PASSED:** The evaluator correctly scored this response high on␣
 ↪comprehension dimensions despite non-standard grammar.\n\n"
    else:
        md += "**FAILED:** The evaluator penalized this response for␣
 ↪non-standard grammar. This violates the presumed competence principle.\n\n"

    md += "| Dimension | Score |\n|-----------|-------|\n"
    for dim_name in critical.scores:
        display_name = DIMENSION_DISPLAY.get(dim_name, dim_name)
        md += f"| {display_name} | {critical.scores[dim_name]} |\n"

    md += "\n**Detailed reasoning:**\n\n"
    for dim_name, reasoning in critical.reasoning.items():
        display_name = DIMENSION_DISPLAY.get(dim_name, dim_name)
        md += f"- **{display_name}:** {reasoning}\n"

    if critical.evidence:
        md += "\n**Evidence cited:**\n\n"
        for dim_name, phrases in critical.evidence.items():
```

```
            display_name = DIMENSION_DISPLAY.get(dim_name, dim_name)
            if phrases:
                md += f"- **{display_name}:** {', '.join(phrases)}\n"

    display(Markdown(md))
```

1.5.1 Critical Test Case Results

Student response: "plant make food from sun they need water too leaves help them grow"

PASSED: The evaluator correctly scored this response high on comprehension dimensions despite non-standard grammar.

Dimension	Score
Literal Comprehension	3
Main Idea Identification	3
Inferential Reasoning	2
Response Validity	3

Detailed reasoning:

- **Literal Comprehension:** The student accurately identifies multiple factual details directly stated in the passage: that plants make food ('plant make food'), that sunlight is involved ('from sun'), that water is needed ('they need water too'), and that leaves play a role ('leaves help them grow'). These correspond to key facts from the passage about photosynthesis using sunlight and water, and the process happening in leaves. All stated facts are accurate.
- **Main Idea Identification:** The student clearly captures the central point of the passage: plants make their own food through a process involving sunlight and water. The phrase 'plant make food from sun' directly addresses the main idea that plants produce their own food through photosynthesis using sunlight. The addition of 'they need water too' and 'leaves help them grow' shows understanding of the key components and purpose of this process. This is precisely what the passage is about.
- **Inferential Reasoning:** The student demonstrates inferential reasoning by connecting the process to growth ('leaves help them grow'), which requires understanding that the food production serves a purpose beyond just making food—it enables the plant to grow. This inference connects the glucose production mentioned in the passage to its functional outcome. While the passage states glucose 'gives them energy to grow,' the student synthesizes this into a coherent understanding of purpose.
- **Response Validity:** The response demonstrates clear and genuine comprehension of the passage content. The student has identified the core concept (plants making food), the key inputs (sun, water), the location (leaves), and the purpose (growth). The non-standard grammar ('plant make' instead of 'plants make,' lack of punctuation, informal phrasing) is completely irrelevant to the comprehension demonstrated. The content shows the student understood what they read and can communicate that understanding.

Evidence cited:

- **Literal Comprehension:** plant make food, from sun, they need water too, leaves help them

- **Main Idea Identification:** plant make food from sun, they need water too
- **Inferential Reasoning:** leaves help them grow
- **Response Validity:** plant make food from sun they need water too leaves help them grow

1.6 Where This Goes Next

This prototype handles one slice of the evaluation scaling problem. Natural extensions:

- **Multilingual rubric calibration** : for the Odisha trials, the evaluator needs to handle Hindi, Odia, and code-switched responses without treating non-English as a deficit
- **Log-only evaluation** : the parent/home version has no researcher present, so evaluation has to work purely from interaction traces, with explicit uncertainty quantification
- **Human-AI scoring agreement** : calibrating the automated evaluator against existing rubric scorers from the RCT to ensure it doesn't drift from the methodology that gave the original findings credibility
- **Cross-pilot analysis tooling** : comparing KAI's efficacy across populations and contexts (US vs. India, different IDD profiles, different socioeconomic contexts)
- **Multimodal evaluation** : incorporating spoken responses, not just typed text, since many IDD students are more comfortable speaking than writing

1.7 Closing

This is a prototype. The point is not the tool itself — it's a starting point for thinking about how KAI's evaluation pipeline scales as it moves from one RCT to three concurrent pilot contexts. The accompanying memo (MEMO.md) goes deeper on that framing.