LLM Specialization and Human-Feedback Evaluation

ISEA Session 12

Dr. Min Sun & Dr. Shawon Sarkar
University of Washington

04.18.2025

IES Institute of Education Sciences

AmplifyLearn.AI

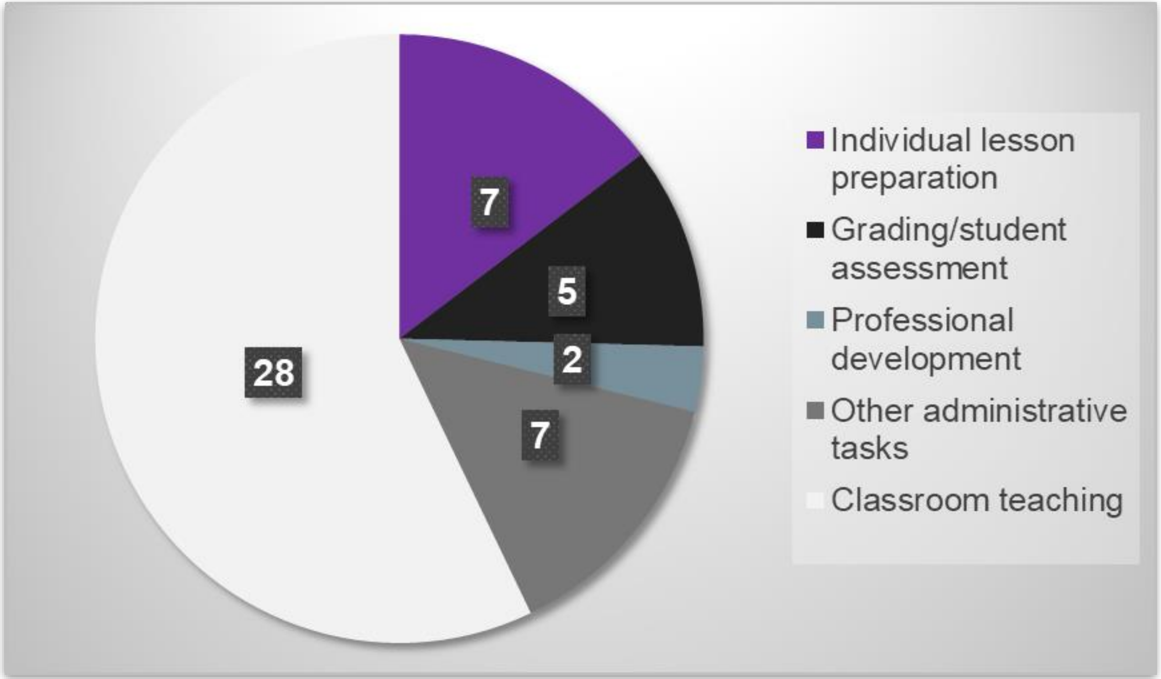
UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Learning Objectives

- Embedding domain knowledge into foundational LLMs
  - Customized prompts
  - Instructional Fine-tuning
- Evaluating your LLM models using human feedback

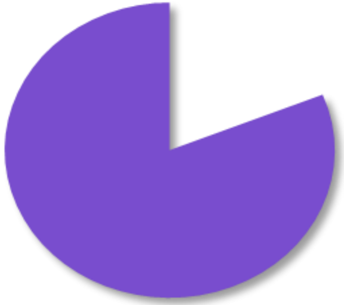
Problem 1: Technology can augment lesson planning to optimize teachers' time allocations



> Teachers spent **7 hours per week** on individual lesson planning and additional **3 more hours** per week if they serve students with diverse learning needs in terms of language, disability, and prior academic learning .

> Although these tasks are essential, they take away teachers' time from the most rewarding part of their profession—engaging students in the classroom.

Problem 2: Foundational LLMs Have Not Solved the Problem, Particularly in Math Education



79%
of a National Sample of Teachers Reported Using ChatGPT in May 2024

Source: Impact Research

Schools are concerned about teachers' overreliance, because ChatGPT:

- Lacks specificity and depth of math tasks,
- Contains mathematical errors,
- Has inadequate understanding of pedagogy and student learning progression,

K-12 education rightfully deserves better technology!

Three Lesson Plan Sources

> Customized GPT 4
> Fine-tuned LLaMA 2-13b

Baseline: Human curriculum designer

Customized GPT 4

- Prompt-engineered GPT-4 with structured, domain-aligned prompts

- Same input template used across all lesson plans generation: subject, grade, title, learning objectives, and CCSSM

- Optimized to generate plans in 4 structured sections: warm-up, main tasks (explain + reinforce), cool-down

Fine-tuned LLaMA 2 13b

- Fine-tuned on open-source math lesson plans curated and revised by educators
- Followed standard training process using supervised fine-tuning

  (Model card) (Meta fine-tuning guide)

  - PEFT, or Parameter Efficient Fine Tuning
  - There are two important PEFT methods: LoRA (Low Rank Adaptation) and QLoRA (Quantized LoRA), where pre-trained models are loaded to GPU as quantized 8-bit and 4-bit weights, respectively.
- Same input template used across all lesson plans generation: subject, grade, title, learning objectives, and CCSSM
- Optimized to generate plans in 4 structured sections: warm-up, main tasks (explain + reinforce), cool-down

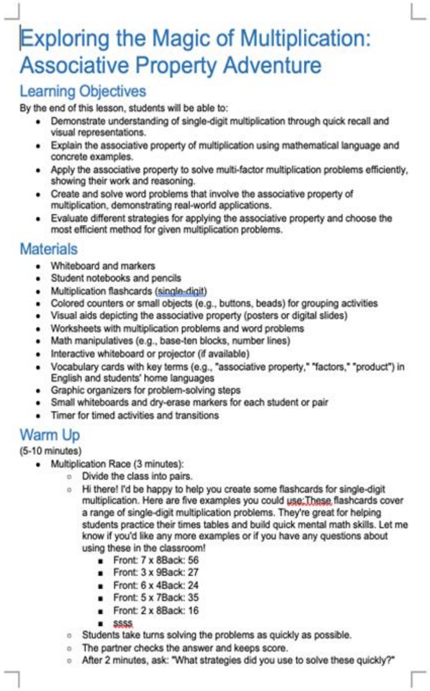
Human Curriculum Designer

- Lesson plans sourced from repositories like Illustrative Mathematics

- Reviewed and selected by researchers for pedagogical quality and grade appropriateness

- Lesson plans filtered using semantic similarity search with text embeddings

Lesson Plan



Exploring the Magic of Multiplication: Associative Property Adventure

Learning Objectives

By the end of this lesson, students will be able to:

- Demonstrate understanding of single-digit multiplication through quick recall and visual representations.
- Explain the associative property of multiplication using mathematical language and concrete examples.
- Apply the associative property to solve multi-factor multiplication problems efficiently, showing their work and reasoning.
- Create and solve word problems that involve the associative property of multiplication, demonstrating real-world applications.
- Evaluate different strategies for applying the associative property and choose the most efficient method for given multiplication problems.

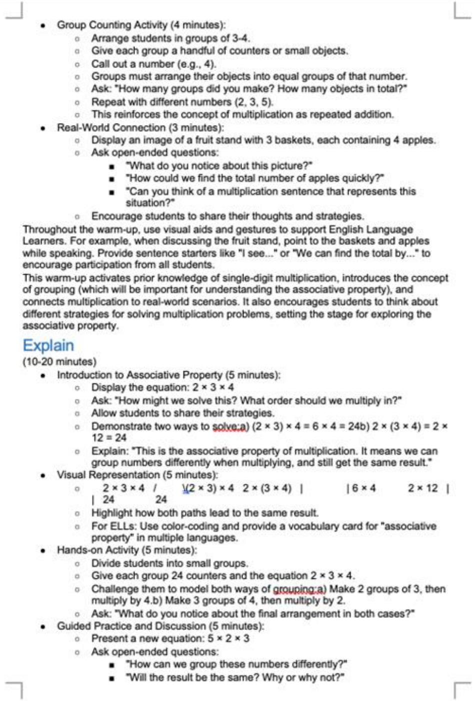
Materials

- Whiteboard and markers
- Student notebooks and pencils
- Multiplication flashcards (single-digit)
- Colored counters or small objects (e.g., buttons, beads) for grouping activities
- Visual aids depicting the associative property (posters or digital slides)
- Worksheets with multiplication problems and word problems
- Math manipulatives (e.g., base-ten blocks, number lines)
- Interactive whiteboard or projector (if available)
- Vocabulary cards with key terms (e.g., "associative property," "factors," "product") in English and students' home languages
- Graphic organizers for problem-solving steps
- Small whiteboards and dry-erase markers for each student or pair
- Timer for timed activities and transitions

Warm Up

(5-10 minutes)

- Multiplication Race (3 minutes):
  - Divide the class into pairs.
  - Hi there! I'd be happy to help you create some flashcards for single-digit multiplication. Here are five examples you could use:These flashcards cover a range of single-digit multiplication problems. They're great for helping students practice their times tables and build quick mental math skills. Let me know if you'd like any more examples or if you have any questions about using these in the classroom!
    - Front: 7 x 8Back: 56
    - Front: 3 x 9Back: 27
    - Front: 6 x 4Back: 24
    - Front: 5 x 7Back: 35
    - Front: 2 x 8Back: 16
    - ssss
  - Students take turns solving the problems as quickly as possible.
  - The partner checks the answer and keeps score.
  - After 2 minutes, ask: "What strategies did you use to solve these quickly?"



- Group Counting Activity (4 minutes):
  - Arrange students in groups of 3-4.
  - Give each group a handful of counters or small objects.
  - Call out a number (e.g., 4).
  - Groups must arrange their objects into equal groups of that number.
  - Ask: "How many groups did you make? How many objects in total?"
  - Repeat with different numbers (2, 3, 5).
  - This reinforces the concept of multiplication as repeated addition.
- Real-World Connection (3 minutes):
  - Display an image of a fruit stand with 3 baskets, each containing 4 apples.
  - Ask open-ended questions:
    - "What do you notice about this picture?"
    - "How could we find the total number of apples quickly?"
    - "Can you think of a multiplication sentence that represents this situation?"
  - Encourage students to share their thoughts and strategies.

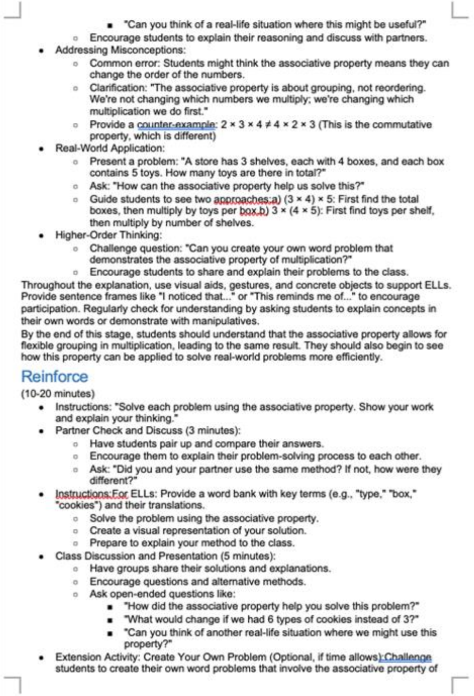
Throughout the warm-up, use visual aids and gestures to support English Language Learners. For example, when discussing the fruit stand, point to the baskets and apples while speaking. Provide sentence starters like "I see..." or "We can find the total by..." to encourage participation from all students.

This warm-up activates prior knowledge of single-digit multiplication, introduces the concept of grouping (which will be important for understanding the associative property), and connects multiplication to real-world scenarios. It also encourages students to think about different strategies for solving multiplication problems, setting the stage for exploring the associative property.

Explain

(10-20 minutes)

- Introduction to Associative Property (5 minutes):
  - Display the equation: 2 × 3 × 4
  - Ask: "How might we solve this? What order should we multiply in?"
  - Allow students to share their strategies.
  - Demonstrate two ways to solve:a) (2 × 3) × 4 = 6 × 4 = 24b) 2 × (3 × 4) = 2 × 12 = 24
  - Explain: "This is the associative property of multiplication. It means we can group numbers differently when multiplying, and still get the same result."
- Visual Representation (5 minutes):
  - 2 × 3 × 4 / \(2 × 3) × 4 2 × (3 × 4) | |6 × 4 2 × 12 | | 24 24
  - Highlight how both paths lead to the same result.
  - For ELLs: Use color-coding and provide a vocabulary card for "associative property" in multiple languages.
- Hands-on Activity (5 minutes):
  - Divide students into small groups.
  - Give each group 24 counters and the equation 2 × 3 × 4.
  - Challenge them to model both ways of grouping:a) Make 2 groups of 3, then multiply by 4.b) Make 3 groups of 4, then multiply by 2.
  - Ask: "What do you notice about the final arrangement in both cases?"
- Guided Practice and Discussion (5 minutes):
  - Present a new equation: 5 × 2 × 3
  - Ask open-ended questions:
    - "How can we group these numbers differently?"
    - "Will the result be the same? Why or why not?"



    - "Can you think of a real-life situation where this might be useful?"
  - Encourage students to explain their reasoning and discuss with partners.
- Addressing Misconceptions:
  - Common error: Students might think the associative property means they can change the order of the numbers.
  - Clarification: "The associative property is about grouping, not reordering. We're not changing which numbers we multiply; we're changing which multiplication we do first."
  - Provide a counter-example: 2 × 3 × 4 ≠ 4 × 2 × 3 (This is the commutative property, which is different)
- Real-World Application:
  - Present a problem: "A store has 3 shelves, each with 4 boxes, and each box contains 5 toys. How many toys are there in total?"
  - Ask: "How can the associative property help us solve this?"
  - Guide students to see two approaches:a) (3 × 4) × 5: First find the total boxes, then multiply by toys per box.b) 3 × (4 × 5): First find toys per shelf, then multiply by number of shelves.
- Higher-Order Thinking:
  - Challenge question: "Can you create your own word problem that demonstrates the associative property of multiplication?"
  - Encourage students to share and explain their problems to the class.

Throughout the explanation, use visual aids, gestures, and concrete objects to support ELLs. Provide sentence frames like "I noticed that..." or "This reminds me of..." to encourage participation. Regularly check for understanding by asking students to explain concepts in their own words or demonstrate with manipulatives.

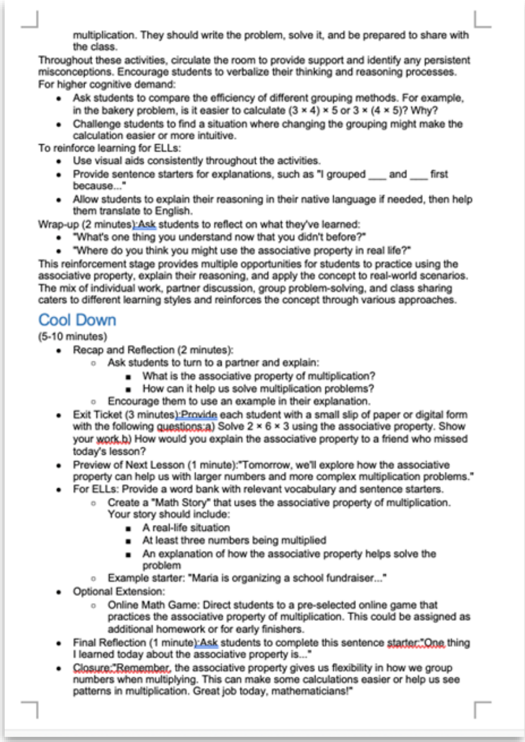
By the end of this stage, students should understand that the associative property allows for flexible grouping in multiplication, leading to the same result. They should also begin to see how this property can be applied to solve real-world problems more efficiently.

Reinforce

(10-20 minutes)

- Instructions: "Solve each problem using the associative property. Show your work and explain your thinking."
- Partner Check and Discuss (3 minutes):
  - Have students pair up and compare their answers.
  - Encourage them to explain their problem-solving process to each other.
  - Ask: "Did you and your partner use the same method? If not, how were they different?"
- Instructions:For ELLs: Provide a word bank with key terms (e.g., "type," "box," "cookies") and their translations.
  - Solve the problem using the associative property.
  - Create a visual representation of your solution.
  - Prepare to explain your method to the class.
- Class Discussion and Presentation (5 minutes):
  - Have groups share their solutions and explanations.
  - Encourage questions and alternative methods.
  - Ask open-ended questions like:
    - "How did the associative property help you solve this problem?"
    - "What would change if we had 6 types of cookies instead of 3?"
    - "Can you think of another real-life situation where we might use this property?"
- Extension Activity: Create Your Own Problem (Optional, if time allows):Challenge students to create their own word problems that involve the associative property of

Lesson Plan (contd.)



multiplication. They should write the problem, solve it, and be prepared to share with the class.

Throughout these activities, circulate the room to provide support and identify any persistent misconceptions. Encourage students to verbalize their thinking and reasoning processes.

For higher cognitive demand:

- Ask students to compare the efficiency of different grouping methods. For example, in the bakery problem, is it easier to calculate (3 × 4) × 5 or 3 × (4 × 5)? Why?
- Challenge students to find a situation where changing the grouping might make the calculation easier or more intuitive.

To reinforce learning for ELLs:

- Use visual aids consistently throughout the activities.
- Provide sentence starters for explanations, such as "I grouped ___ and ___ first because..."
- Allow students to explain their reasoning in their native language if needed, then help them translate to English.

Wrap-up (2 minutes):Ask students to reflect on what they've learned:

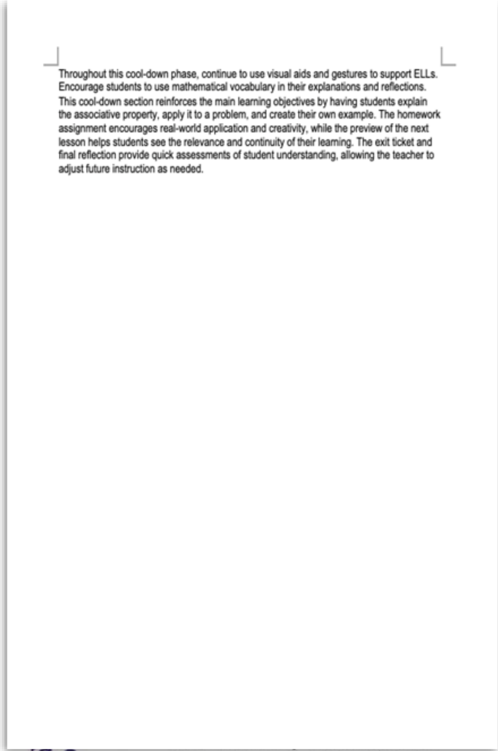
- "What's one thing you understand now that you didn't before?"
- "Where do you think you might use the associative property in real life?"

This reinforcement stage provides multiple opportunities for students to practice using the associative property, explain their reasoning, and apply the concept to real-world scenarios. The mix of individual work, partner discussion, group problem-solving, and class sharing caters to different learning styles and reinforces the concept through various approaches.

Cool Down

(5-10 minutes)

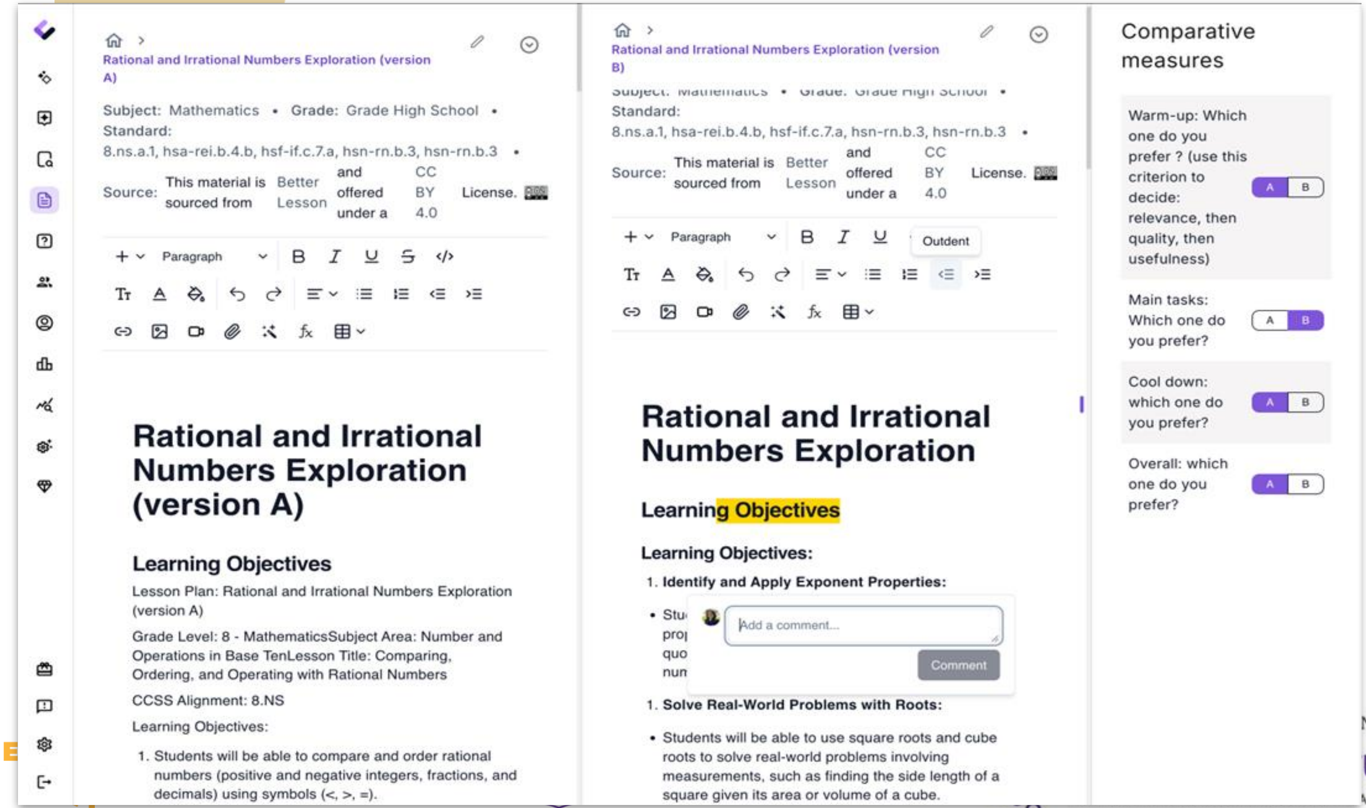
- Recap and Reflection (2 minutes):
  - Ask students to turn to a partner and explain:
    - What is the associative property of multiplication?
    - How can it help us solve multiplication problems?
  - Encourage them to use an example in their explanation.
- Exit Ticket (3 minutes):Provide each student with a small slip of paper or digital form with the following questions:a) Solve 2 × 6 × 3 using the associative property. Show your work.b) How would you explain the associative property to a friend who missed today's lesson?
- Preview of Next Lesson (1 minute):"Tomorrow, we'll explore how the associative property can help us with larger numbers and more complex multiplication problems."
- For ELLs: Provide a word bank with relevant vocabulary and sentence starters.
  - Create a "Math Story" that uses the associative property of multiplication. Your story should include:
    - A real-life situation
    - At least three numbers being multiplied
    - An explanation of how the associative property helps solve the problem
  - Example starter: "Maria is organizing a school fundraiser..."
- Optional Extension:
  - Online Math Game: Direct students to a pre-selected online game that practices the associative property of multiplication. This could be assigned as additional homework or for early finishers.
- Final Reflection (1 minute):Ask students to complete this sentence starter:"One thing I learned today about the associative property is..."
- Closure:"Remember, the associative property gives us flexibility in how we group numbers when multiplying. This can make some calculations easier or help us see patterns in multiplication. Great job today, mathematicians!"



Throughout this cool-down phase, continue to use visual aids and gestures to support ELLs. Encourage students to use mathematical vocabulary in their explanations and reflections.

This cool-down section reinforces the main learning objectives by having students explain the associative property, apply it to a problem, and create their own example. The homework assignment encourages real-world application and creativity, while the preview of the next lesson helps students see the relevance and continuity of their learning. The exit ticket and final reflection provide quick assessments of student understanding, allowing the teacher to adjust future instruction as needed.

Study Design



Rational and Irrational Numbers Exploration (version A)

Subject: Mathematics • Grade: Grade High School • Standard: 8.ns.a.1, hsa-rei.b.4.b, hsf-if.c.7.a, hsn-rn.b.3, hsn-rn.b.3 •

Source: This material is sourced from Better Lesson and offered under a CC BY 4.0 License.

Rational and Irrational Numbers Exploration (version A)

Learning Objectives

Lesson Plan: Rational and Irrational Numbers Exploration (version A)

Grade Level: 8 - MathematicsSubject Area: Number and Operations in Base TenLesson Title: Comparing, Ordering, and Operating with Rational Numbers

CCSS Alignment: 8.NS

Learning Objectives:

1. Students will be able to compare and order rational numbers (positive and negative integers, fractions, and decimals) using symbols (<, >, =).

Rational and Irrational Numbers Exploration (version B)

Subject: Mathematics • Grade: Grade High School • Standard: 8.ns.a.1, hsa-rei.b.4.b, hsf-if.c.7.a, hsn-rn.b.3, hsn-rn.b.3 •

Source: This material is sourced from Better Lesson and offered under a CC BY 4.0 License.

Rational and Irrational Numbers Exploration

Learning Objectives

Learning Objectives:

1. Identify and Apply Exponent Properties:
   - Stu prop quo nun

1. Solve Real-World Problems with Roots:
   - Students will be able to use square roots and cube roots to solve real-world problems involving measurements, such as finding the side length of a square given its area or volume of a cube.

Comparative measures

Warm-up: Which one do you prefer ? (use this criterion to decide: relevance, then quality, then usefulness) [✓ A] B

Main tasks: Which one do you prefer? A [✓ B]

Cool down: which one do you prefer? [✓ A] B

Overall: which one do you prefer? [✓ A] B

20 experienced math teachers evaluated model performance and collect data for human feedback reinforcement learning to refine the model.

Binary preference rating

UNIVERSITY OF OREGON

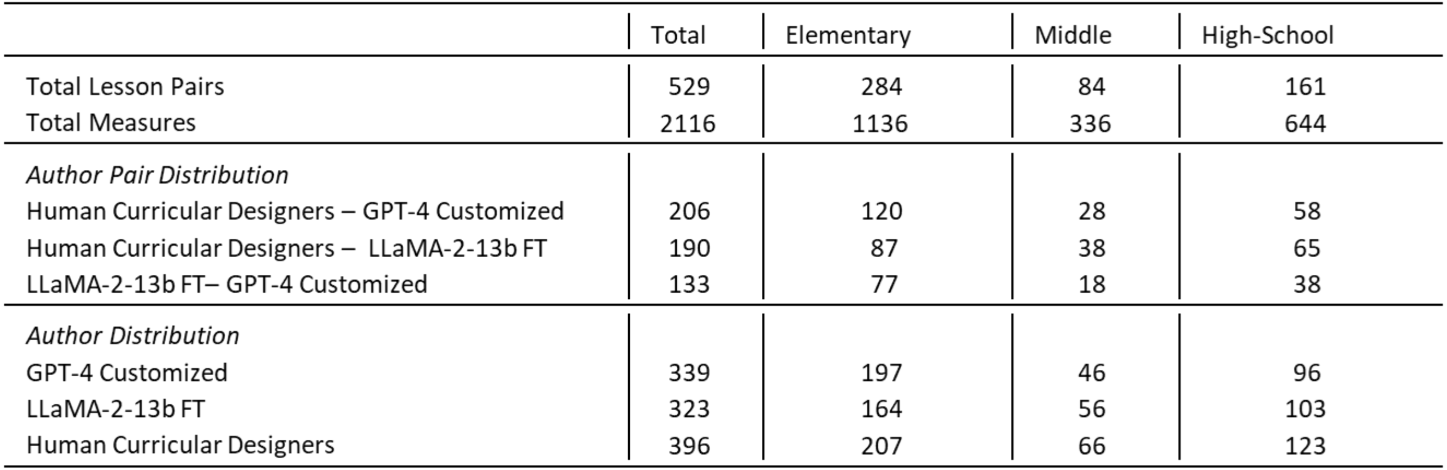
LLM Evaluation with Human Feedback

- Human preference served as evaluation signal across structured lesson plan components
- Educators assessed lesson quality: warm-up, main tasks, cool-down, and overall
- Feedback collected via binary ratings and open-ended comments
- Enabled evaluation of: Prompted vs. fine-tuned model performance
- Grade-level and section-level alignment with educator expectations
- Pedagogical coherence, rigor, and support for diverse learners

Data Collection

Three authors: human curricular designers (Illustrative Math), GPT4 customized, LLaMA-2-13b fine-tuned (FT)

Table 1. Dataset Description



	Total	Elementary	Middle	High-School
Total Lesson Pairs	529	284	84	161
Total Measures	2116	1136	336	644
Author Pair Distribution				
Human Curricular Designers – GPT-4 Customized	206	120	28	58
Human Curricular Designers – LLaMA-2-13b FT	190	87	38	65
LLaMA-2-13b FT– GPT-4 Customized	133	77	18	38
Author Distribution				
GPT-4 Customized	339	197	46	96
LLaMA-2-13b FT	323	164	56	103
Human Curricular Designers	396	207	66	123

IES Institute of Education Sciences W AmplifyLearn.AI UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS O UNIVERSITY OF OREGON

Activity: Breakout Room

In groups of 3-4, examine and evaluate AI-generated educational content through the lens of your professional expertise

> Group Task: Analyze sample lesson plans linked

Lesson Plan 1, Lesson Plan 2

> Critically reflect on either:

- LLM specialization strategy
- the design of this study

IES Institute of Education Sciences W AmplifyLearn.AI UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS O UNIVERSITY OF OREGON

Activity: Breakout Room

Discussion Prompts:

LLM specialization

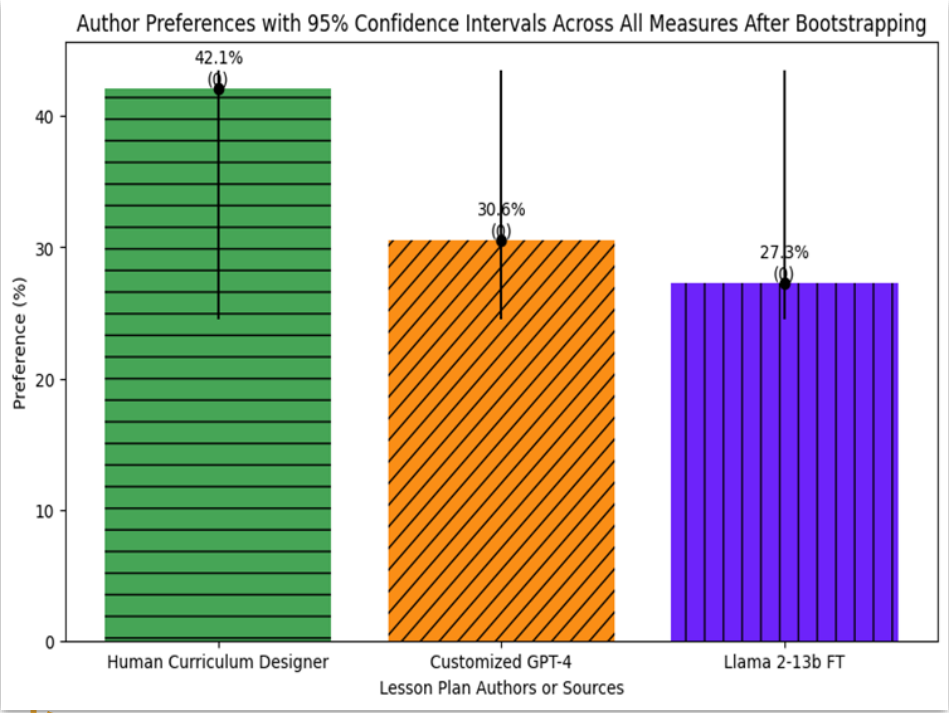
> Any questions about the basic logic of system prompt specialization

> Any questions about basic flow of fine-tuning

Human-feedback evaluation:

- What kinds of educator feedback should future LLM evaluation pipelines capture?
- In your role, what criteria do you currently use to evaluate the effectiveness of AI generated resources or tools?

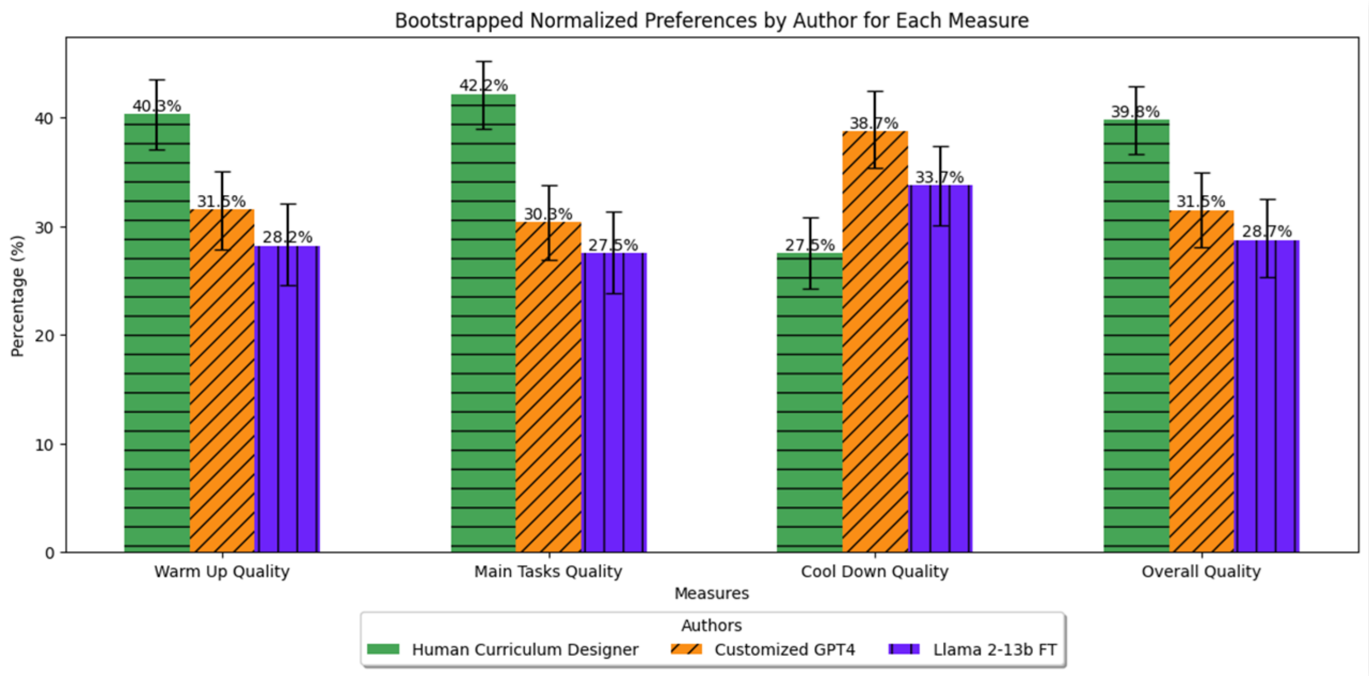
Result 1: Overall Model Performance by Authors



Human created content is overall preferred. Raw differences is about ~12% between human and customized GPT-4 and ~15% with LLaMA 2-13b fine-tuned model (FT).

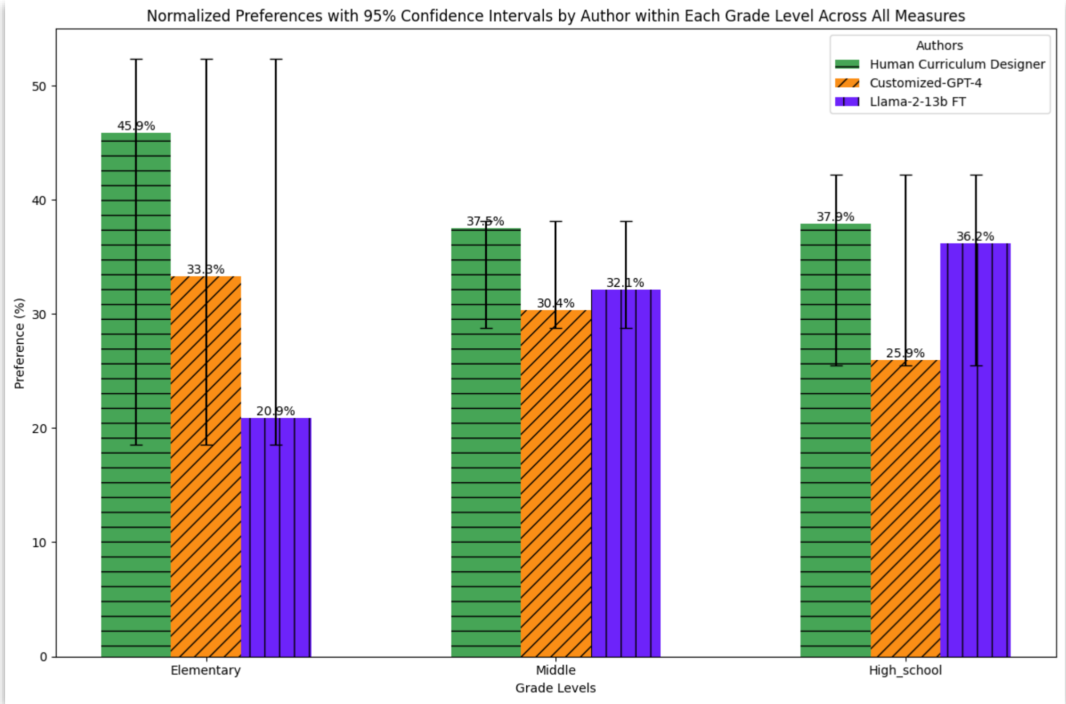
IES Institute of Education Sciences
W AmplifyLearn.AI
UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS
O UNIVERSITY OF OREGON

Result 2: Overall Performance by Authors and Measures



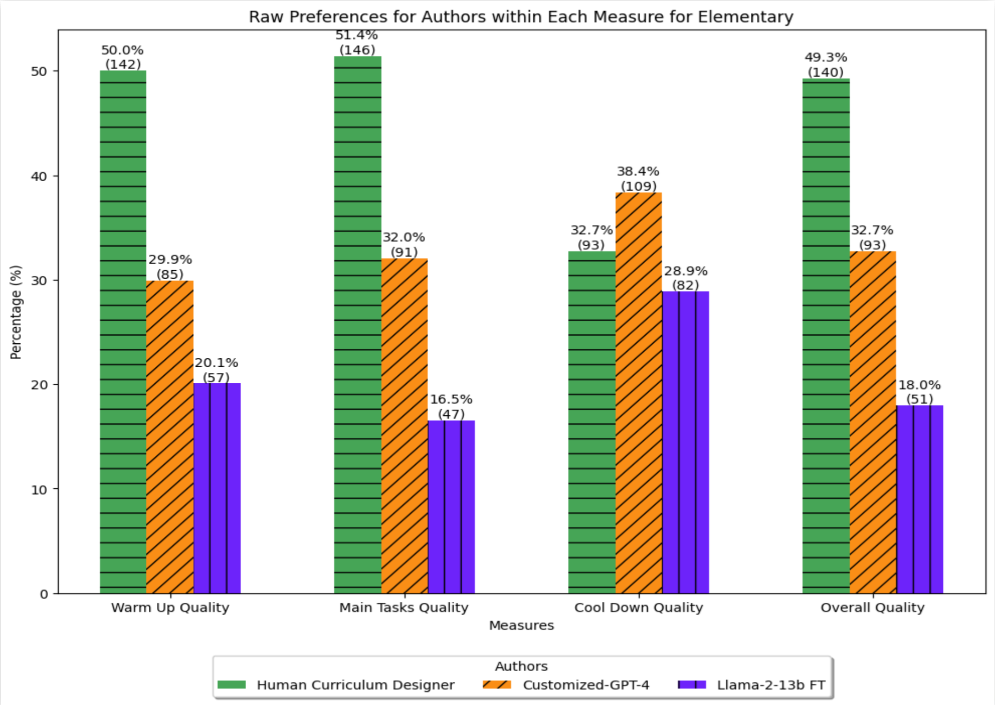
Although human-created content is preferred overall, AI generated cool-down section is preferred on average.

Result 3. Overall Performance By Educational Levels



The preference gap mainly exists in elementary (k-5) lesson plans, while the gaps in secondary levels (middle and high school) are significantly smaller.

Result 4. Overall Performance at Elementary Level By Measures



At elementary level, human-created warm up, main tasks, and overall quality are preferred by experience teachers, while AI-generated cool down is preferred.

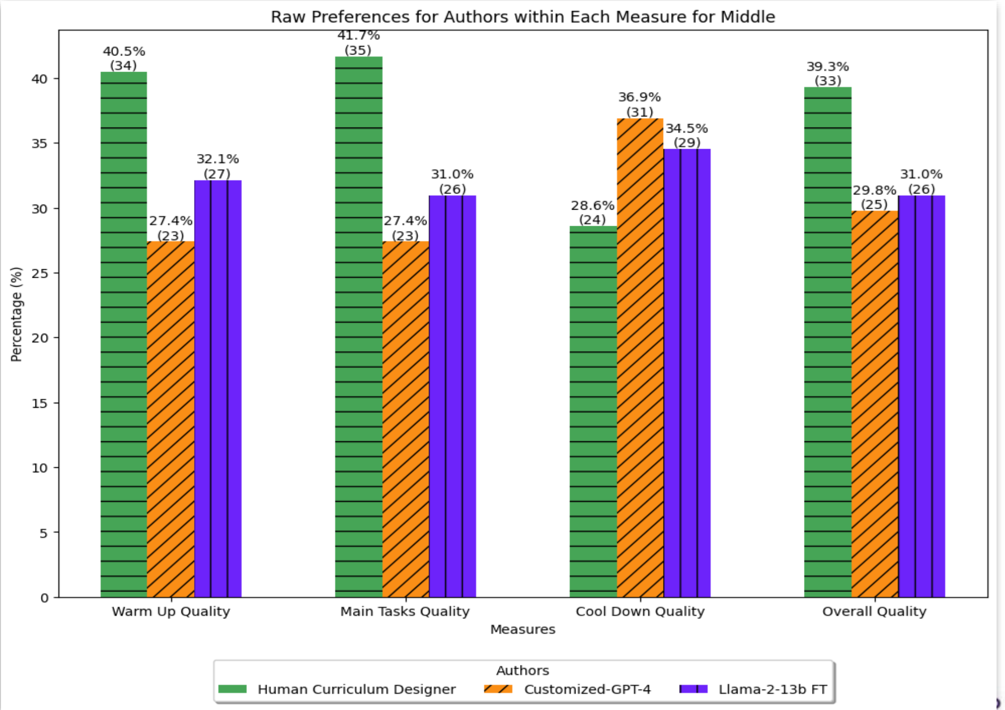
IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

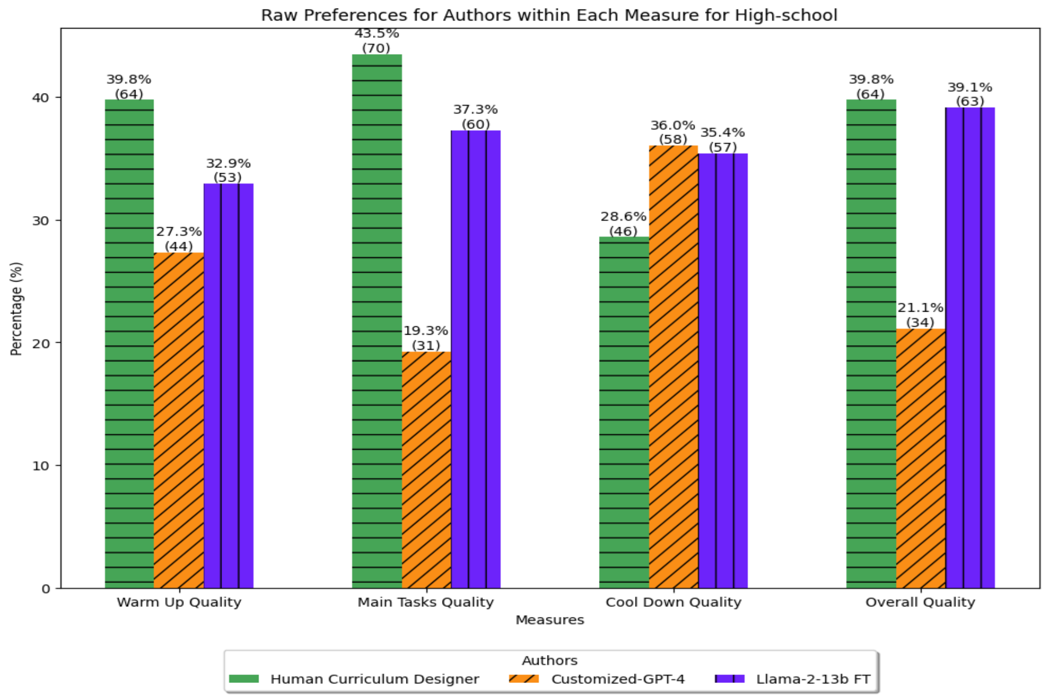
UNIVERSITY OF OREGON

Result 5. Overall Performance at Middle Grade Level By Measures



At middle grade level, in addition to outperformed AI-generated cool down, there are smaller preference gaps between AI-generated and human created in other measures too, The two AI authors are largely at a similar level of performance.

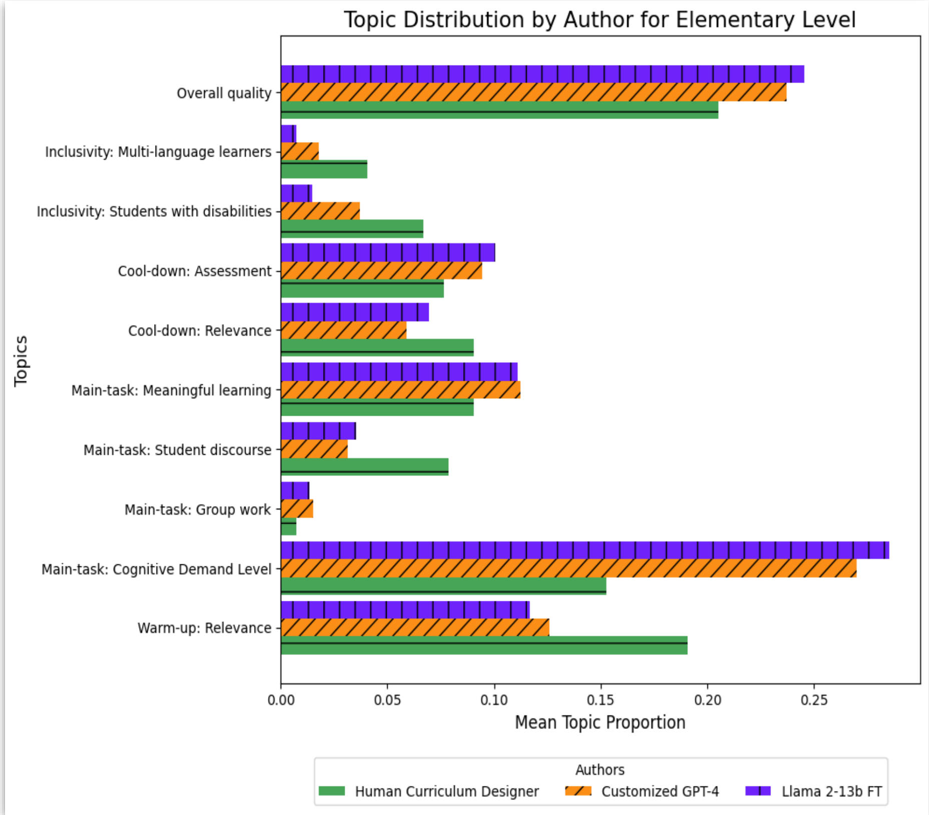
Result 5. Overall Performance at High School Level By Measures



At high school level, a noticeable pattern is that Llama FT outperformed GPT-4 on warm up and main task quality, on par on cool down, thus, significantly outperformed on overall quality than GPT-4 and on par with human-created lessons.

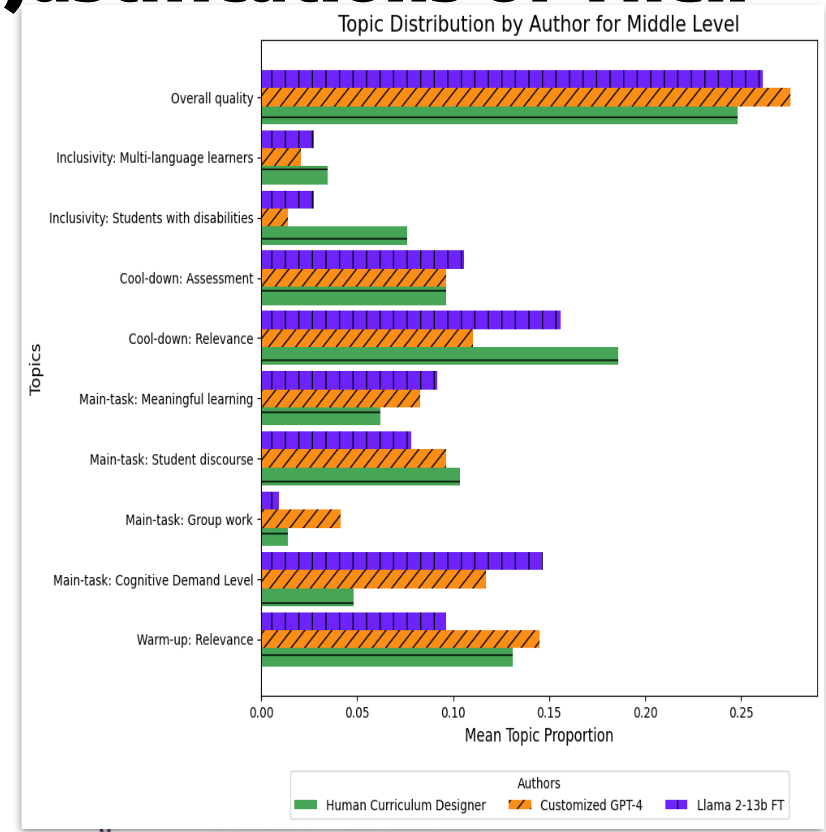
Teachers' Comments and Justifications of Their Choice

<span style="color:purple">Elementary</span> teachers mainly commented on overall quality, main-task cognitive demand level, meaningful learning, warm-up relevance and cool-own assessment and relevance.



Teachers’ Comments and Justifications of Their Choice

Middle grade teachers mainly commented on overall quality, cool down, then cognitive demand level.



IES Institute of Education Sciences

AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

UNIVERSITY OF OREGON

Teachers' Comments and Justifications of Their Choice

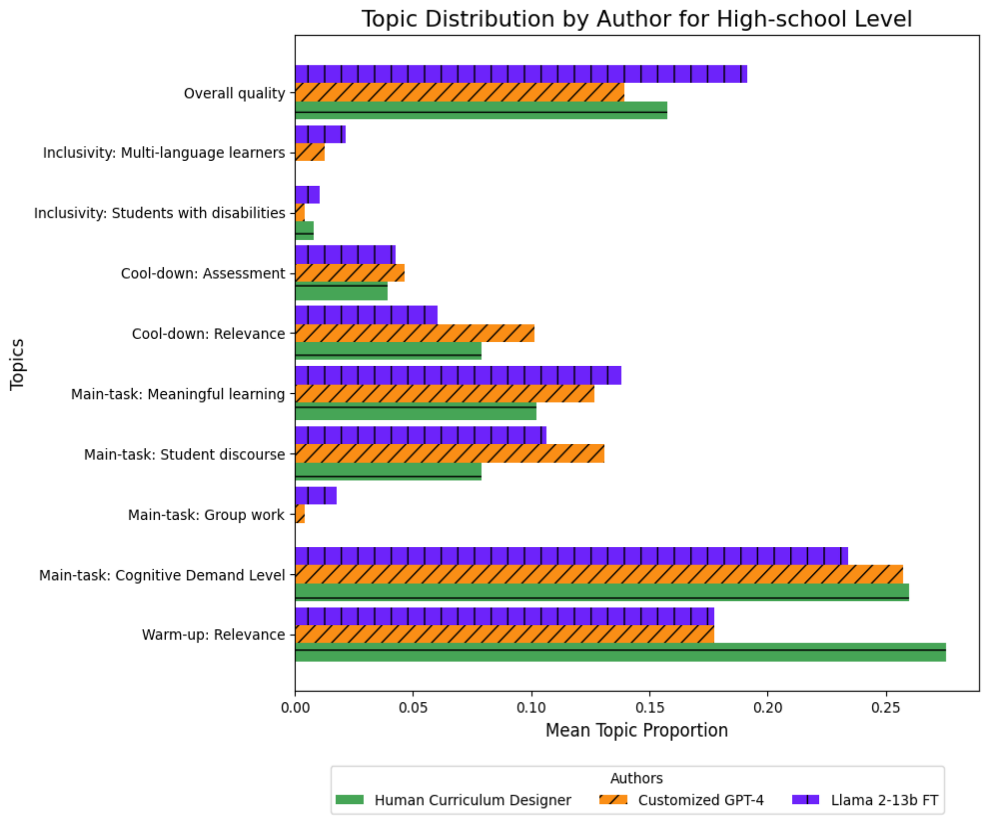
High school teachers mainly commented on main-task cognitive demand level, warm-up relevance, overall quality main task: meaningful learning, and student discourse.

Topic Modeling (LDA)

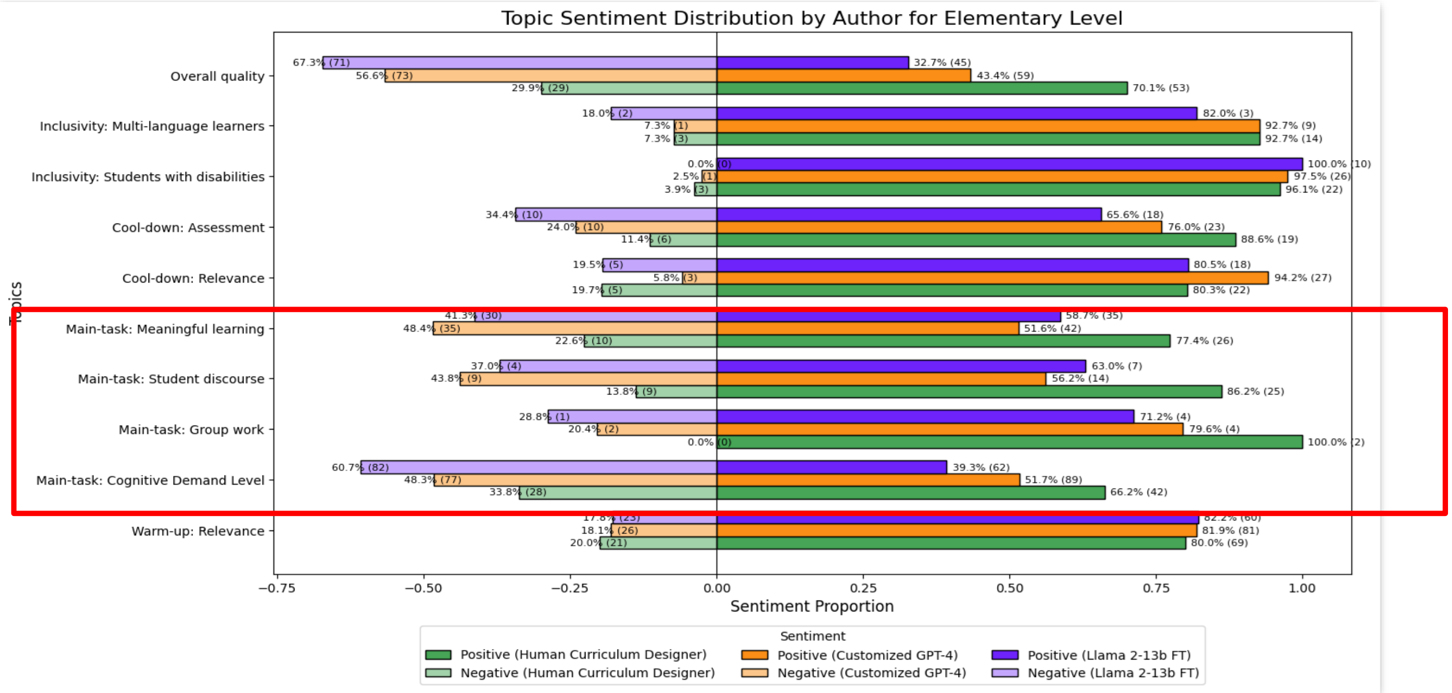
Applied LDA on educator feedback comments

Extracted key themes: student discourse, scaffolded support, group work, etc.

Validated topics using manual thematic coding



Teachers Justifications by Sentiment



At Elementary school level, teachers consistently praised human created lessons having higher quality of main task: cognitive demand level, meaningful learning activities, and effective group work, as well as cool down assessment

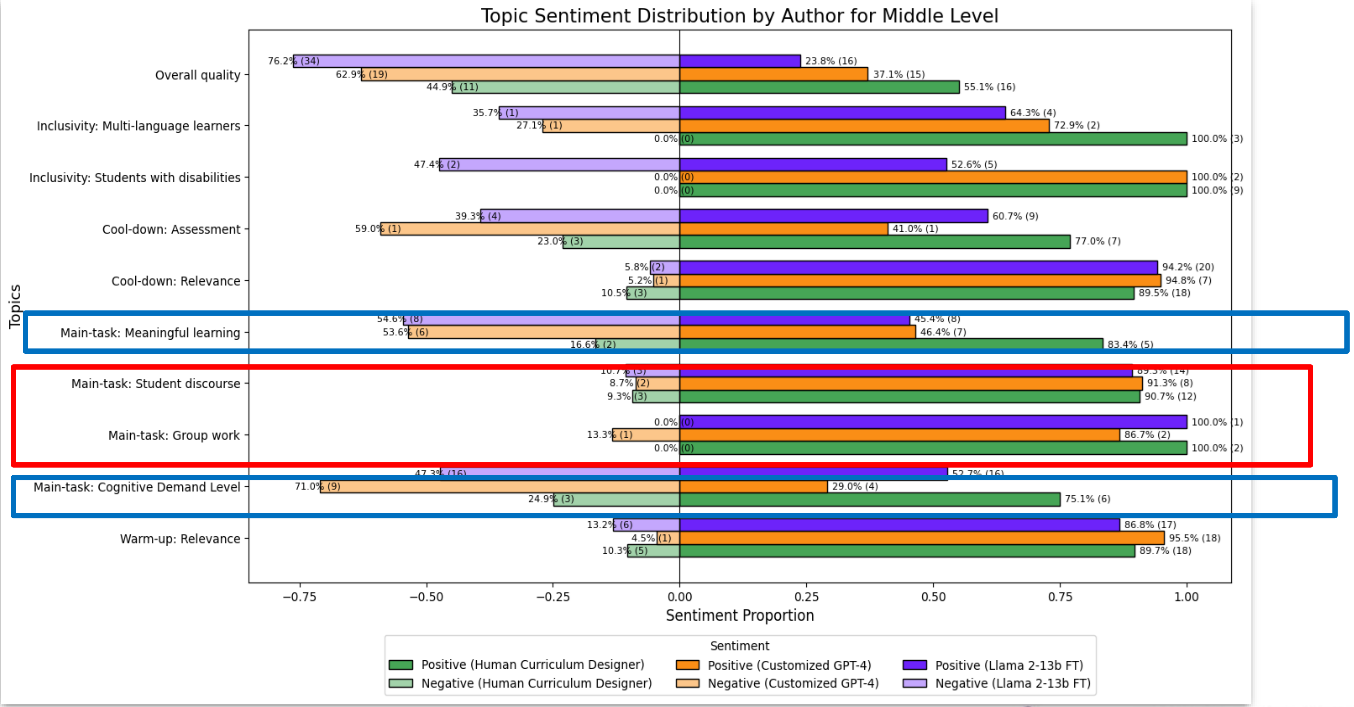
IES Institute of Education Sciences

W AmplifyLearn.AI

UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

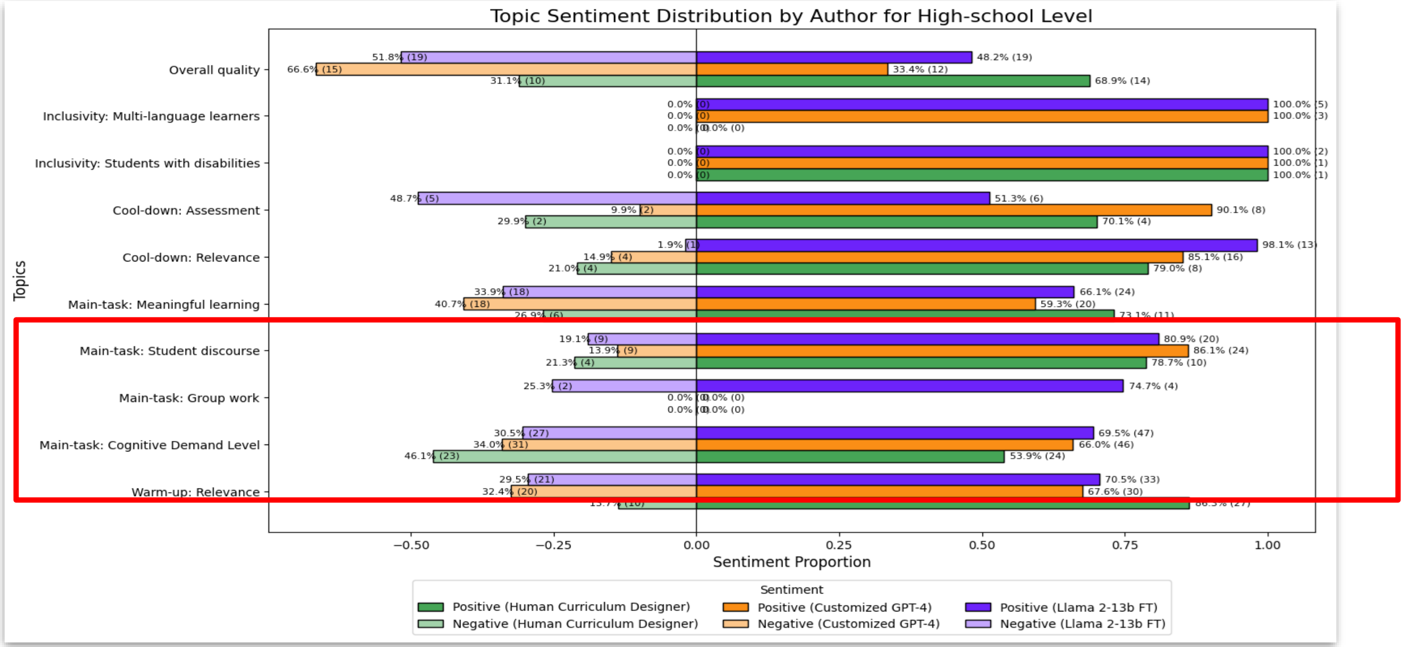
O UNIVERSITY OF OREGON

Teachers Justifications by Sentiment



At middle school level, teachers rated the main task- meaningful learning, group work at the similar level as human-created ones, although they still preferred human-created main task cognitive demand level.

Teachers' Justifications by Sentiment



At high school level, surprisingly, teachers commented more positively on the AI-generated main task cognitive demand level even than human created ones, on par with human-created main task student discourse and meaningful learning. Customized GPT-4 generated higher quality cool down assessment than the other two authors.

Summary of Key Takeaways

- Overall, teachers preferred math lessons created by human curriculum designers.
- However, even with a relatively low effort of specializing AI models using domain knowledge and labelled data, AI is able to generate lesson materials that are better or on par with human-created ones.
    - Cool down section
    - Secondary level, particularly high school
- LlaMa FT's higher performance for high school lessons suggests that for highly domain-specific tasks, instructional fine-tuning with high-quality data can be a cost-effective approach to improve model performance.

IES Institute of Education Sciences W AmplifyLearn.AI UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS O UNIVERSITY OF OREGON

Hackweek Project Proposal

- Propose your own Project for the Hackweek
- Make the proposal meets the hackweek criteria: https://www.amplifylearn.ai/isea/isea-summer-hackweek/
- Fill out the form: Application

No Assignment

- Catch up with the past assignments
- They will give you hands-on practice that will come useful during hackweek