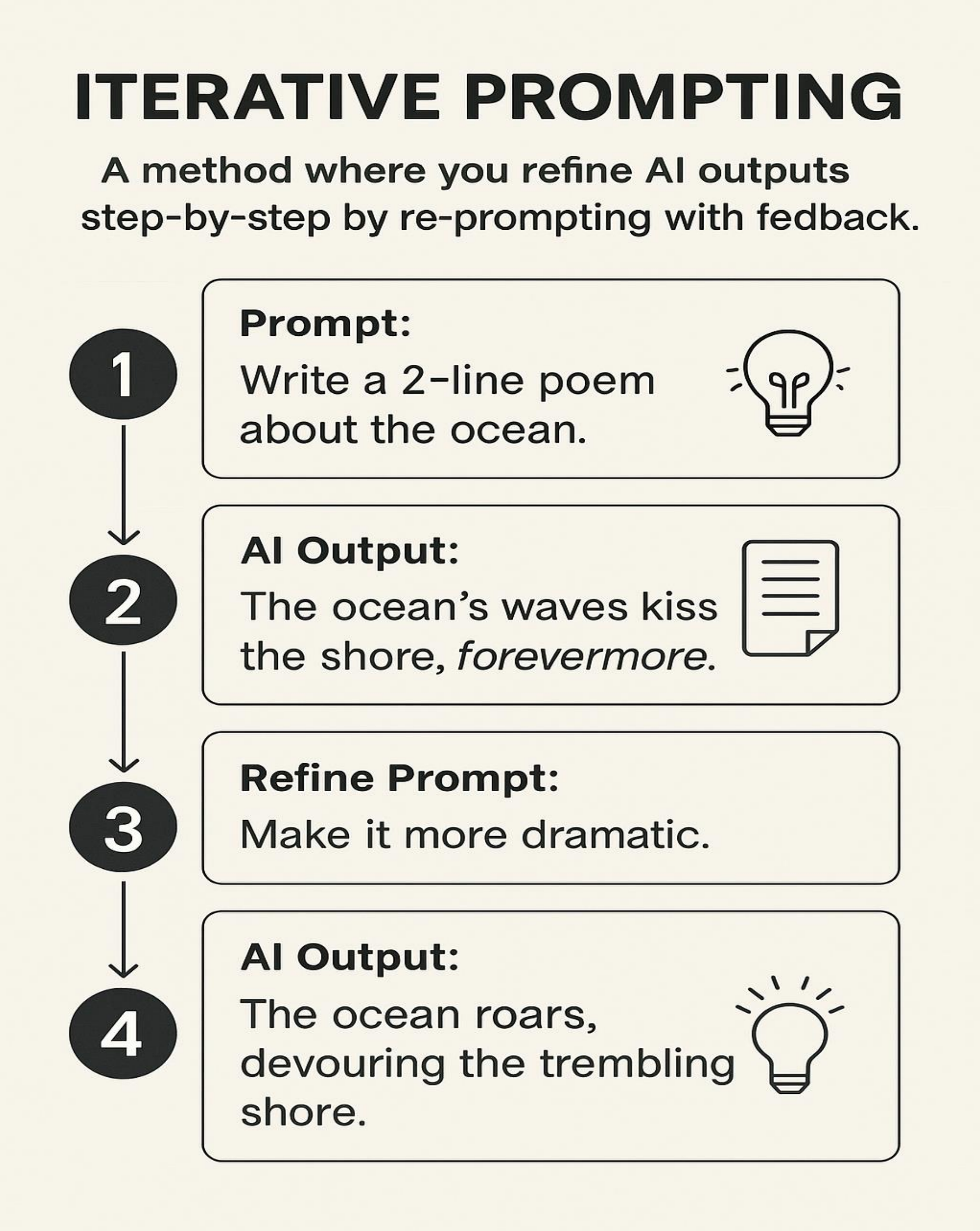
From Outline to Comprehensive Syllabus: Evaluating Consistency in Iterative Prompting

Vickey Ghimire, Bijay Dhungana, Jaljala Shrestha Lama, Nazmus Sadat (Advisor), and Nicholas Caporusso (Advisor)
School of Computing and Analytics, College of Informatics, Northern Kentucky University

What is Iterative Prompting?

Iterative prompting involves using an AI assistant to progressively refine a course outline. It starts with an initial draft, followed by multiple rounds of revisions based on specific instructions, such as "add additional topics" or "enhance detail levels." This method supports the development of comprehensive outlines.

However, a key challenge emerges: the AI frequently generates repeated or rephrased content rather than novel material, increasing the need for manual review. Our research addresses this issue by developing a solution to detect and prioritize genuinely new content.



ITERATIVE PROMPTING

A method where you refine AI outputs step-by-step by re-prompting with fedback.

1 **Prompt:** Write a 2-line poem about the ocean.

2 **AI Output:** The ocean's waves kiss the shore, *forevermore.*

3 **Refine Prompt:** Make it more dramatic.

4 **AI Output:** The ocean roars, devouring the trembling shore.

Fig. 1: Iterative prompting flowchart.

The Challenge: Redundancy in Iteration

To create detailed course outlines, instructors often use iterative prompting—repeatedly asking an LLM to refine and expand its output. However, this process frequently leads to redundant or simply rephrased content, making it difficult to distinguish genuinely new ideas from existing ones. Manually reviewing each iteration is time-consuming and negates many of the efficiency gains that LLMs promise.

Our Solution: A Novelty Metric

A framework was developed that automatically measures the Novelty of content generated by LLMs during iterative prompting.

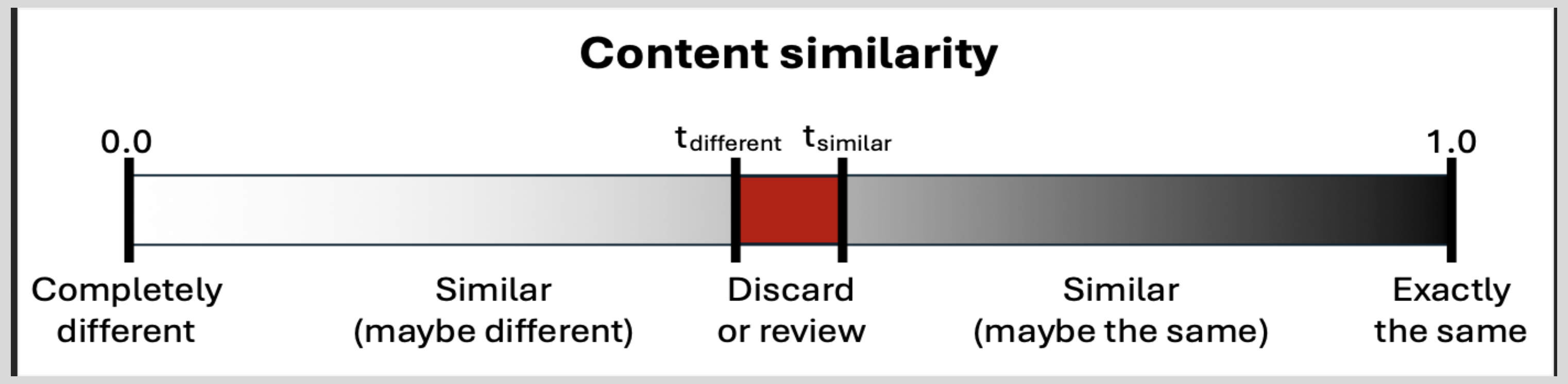


Fig. 2: Content similarity spectrum highlighting the area usually requiring the most human supervision to disambiguate similarity.

Study Protocol

Our study evaluated five leading LLMs: ChatGPT 4o, Claude 3.7 Sonnet, Gemini 2.5 Flash, DeepSeek v3, and DeepSeek r1. The task was to generate a 14-week course outline for Java programming. The process involved:

- An initial detailed prompt to create the course outline.
- Four subsequent prompts asking the LLM to revise and expand upon its previous response.
- This protocol was repeated five times for each model, resulting in 25 conversations and a dataset of over 5,200 individual course topics.

The "Novelty" (NOV) Score Framework

Our NOV score uses a multi-stage process to automatically classify content with high accuracy and low computational cost:

1. **Initial Filter:** We first apply a lightweight lexical method (Jaccard similarity) to quickly identify topics that are "exactly the same" as previous outputs.
2. **Semantic Analysis:** For the remaining content, we use more advanced transformer-based models to measure semantic similarity. This helps distinguish between paraphrased content and genuinely new ideas.
3. **Thresholding:** Using optimized thresholds ($t_{different}$= 0.46 and $t_{similar}$= 0.54), the algorithm classifies ambiguous topics as either "mostly different" (novel) or "mostly similar" (repeated). Content that falls between these thresholds is flagged for manual review.

This hybrid approach solves an optimization problem that maximizes accurate classification while minimizing the number of items that require human intervention.

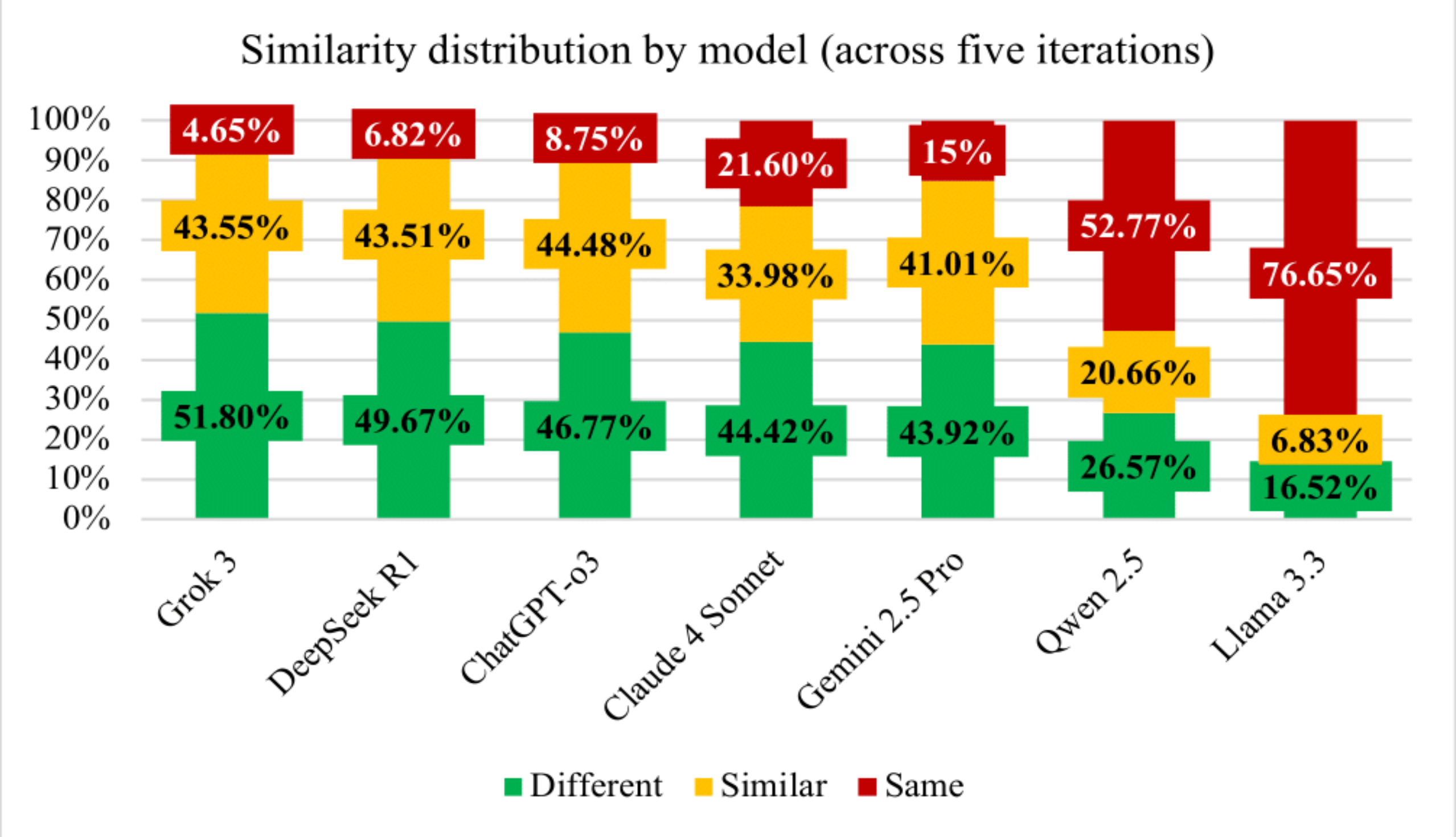
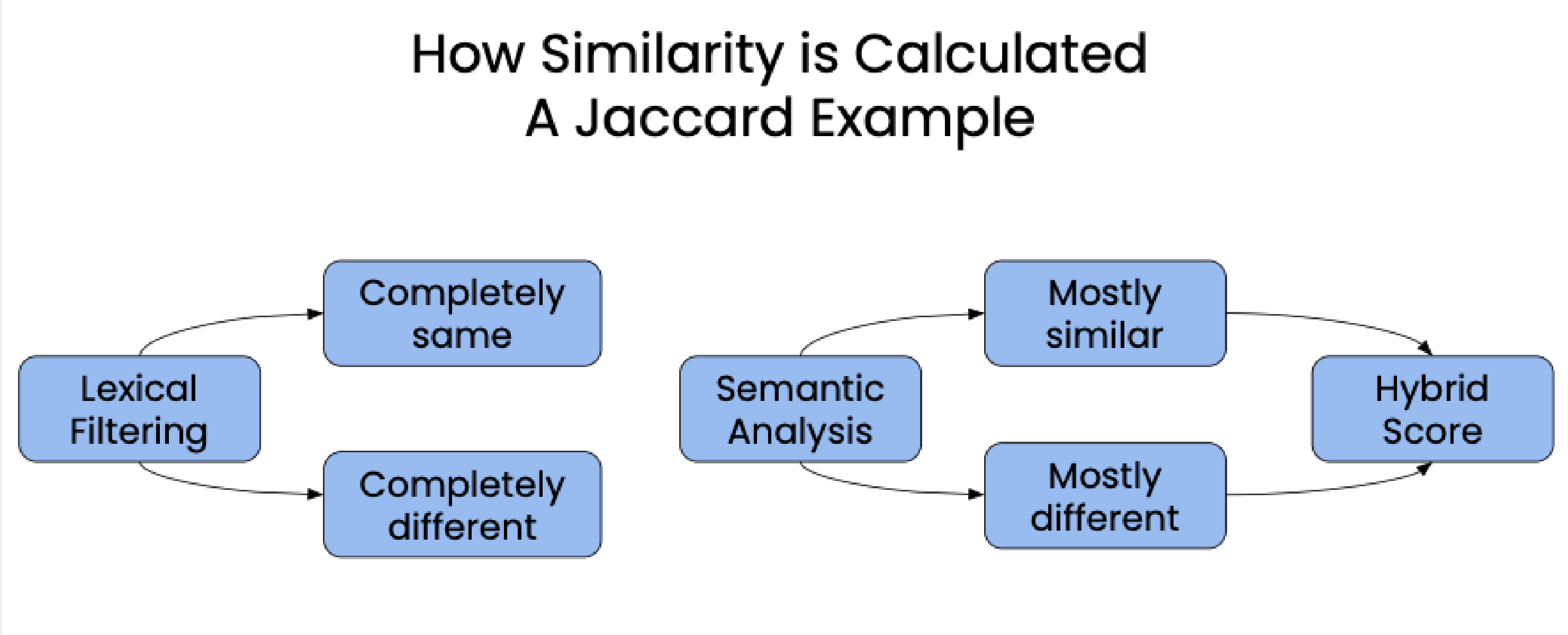


Fig. 3: Initial content similarity distribution by model. Across all models, a significant portion of content (average of 32-47% for top 3 models) was initially classified as "Similar," requiring further analysis.

An Example of How Similarity is Calculated



How Similarity is Calculated
A Jaccard Example

Fig. 4: Jaccard similarity calculation.

Automating Content Classification

The NOV score proved highly effective at reducing the need for manual review.

- Initially, 2,288 topics (**25.28%** of the total) were considered "similar" and would have required human intervention.
- Our NOV score automatically classified **90.52%** of these ambiguous topics, correctly identifying **65.2%** as "mostly different" and **25.31%** as "mostly similar".
- This left only **9.48%** of the ambiguous content needing manual review, streamlining the course creation workflow significantly.

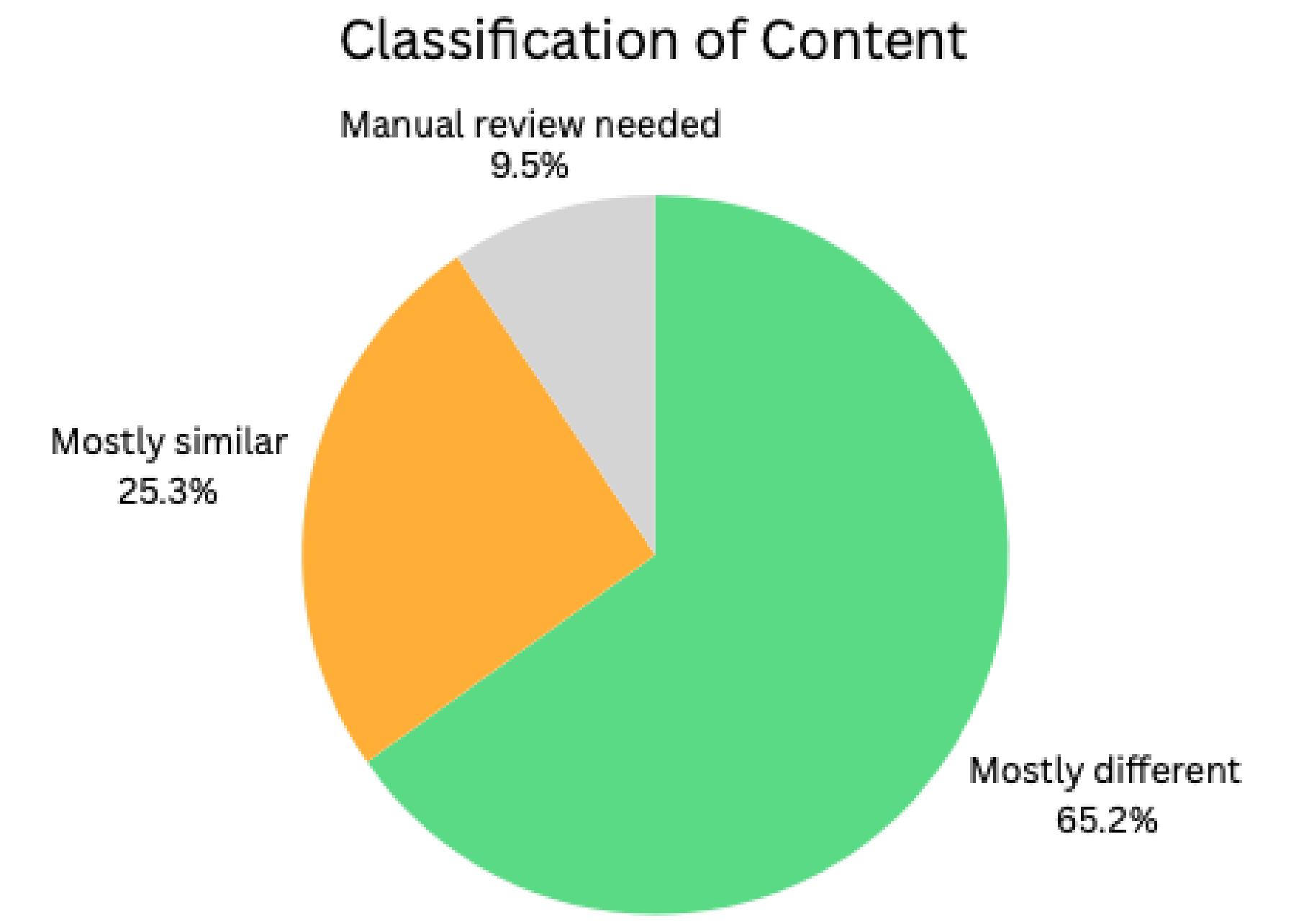


Fig. 5: Breakdown of content classification: manual review vs. automatic classification.

Model Performance Rankings

The NOV score provides a clear benchmark for evaluating how well LLMs introduce new content during iterative prompting.

•**DeepSeek R1** was the top performer, producing **86.28%** novel content across iterations, clearly distancing it from competitors.

•**Qwen 2.5 (42.6%)** and **Llama 3.3 (19.37)** showed high redundancy and were less effective for this task.

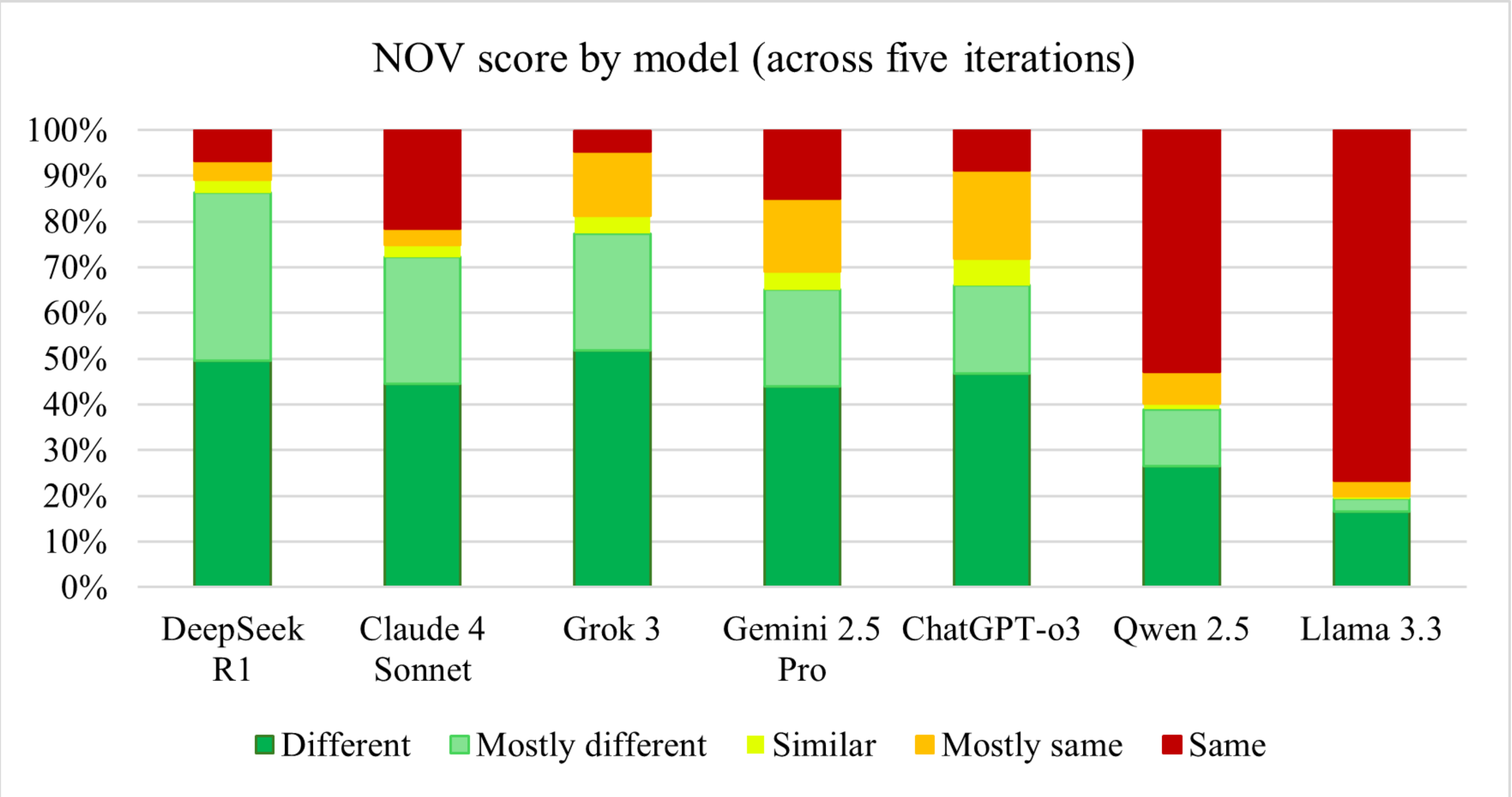


Fig. 6: Final NOV score by model.

Result and Conclusion

When comparing models, DeepSeek r1 stood out by creating the most new and unique content — over 94% was different from before. Other models like Gemini Flash 2.0 and ChatGPT 4o did okay, but some, like Claude 3.7 and DeepSeek v3, repeated themselves a lot. Overall, this shows that our NOV score helps both speed up review and pick the most creative AI models.