**4th Place Solution**

First, I would like to express my gratitude to the host for organizing such an incredible competition. I also want to thank my teammates, [charmq](https://www.kaggle.com/charmq" \t "_blank) and [kami](https://www.kaggle.com/kami634) for their collaboration.

**Summary**

* **Approach to the Competition**
  + Since most misconceptions in the test set were anticipated to be absent in the train set, we focused on improving the accuracy for these unseen misconceptions.
* **Solution Overview**
  + **Data Generation**
    - Data was generated using Qwen2.5-72B-Instruct-AWQ.
    - The process specifically targeted misconceptions not present in the training dataset.
  + **Misconception Generation**
    - Used Qwen2.5-32B-instruct-AWQ to generate misconceptions from Questions, Answers, etc.
  + **Retriever**
    - Training
      * Created two models and fine-tuned them using LoRA.
        + Qwen2.5-14B-instruct
        + Qwen2.5-32B-instruct
      * During training, in addition to the provided data such as QuestionText and AnswerText, misconceptions generated in the previous misconception generation phase were also added to the text.
    - Inference & Retrieval:
      * Conducted retrieval by concatenating the outputs of Qwen2.5-14B-instruct and Qwen2.5-32B-instruct. The outputs were generated for several folds.
      * For each Question-Answer pair, retrieved the following numbers of misconceptions:
        + 25 most similar misconceptions from all misconceptions.
        + 15 most similar misconceptions specifically from those not present in the train set.
        + Removed duplicates.
  + **Reranker**
    - **Training**
      * Fine-tuned multiple Qwen2.5-32B-instruct models using LoRA, adjusting the number of negative samples and adding generated data to the training dataset.
    - **Inferene**
      * Ensembled the LoRA components of the trained models to create three models.
      * Finally, ensembled the outputs of these three models.
  + **Post-Processing**
    - Adjusted the predictions for misconceptions that existed in the train set by reducing their scores:
      * Prediction Score \* 0.40
  + During data generation and inference (e.g., Retreival、Rerank), processes were accelerated using vLLM

**Solution**

**Validation Strategy**

* Group KFold (group = QuestionId)
* 5 folds
* Evaluated not only the overall fold scores but also the scores for validation data extracted specifically for misconceptions not included in the training data.

**Data Generation**

* Used Qwen2.5-72B-Instruct-AWQ to generate new questions, answer choices, and incorrect answers corresponding to each misconception that does not appear in the training dataset.
* Approximately 8000 questions were created, and 2500 of them were randomly sampled for training the reranker model.
* For each misconception, 100 randomly sampled questions from the training dataset were added to the prompt as input examples for data generation.

The prompt for data generation is as follows, with the maximum token count being approximately 20000.

"""You are an expert in mathematics.

Refer to the examples below to create new problem with given misconception.

Misconception: {MisconceptionText}

The output format shoud be below.

ConstructName:

SubjectName:

Math problem:

Answer A text:

Answer B text:

Answer C text:

Answer D text:

Answer:

Incorrect answer:

The examples are below

Example 1:

ConstructName: {ConstructName\_1}

SubjectName: {SubjectName\_1}

Math problem: {QuestionText\_1}

Answer A text: {AnswerAText\_1}

Answer B text: {AnswerBText\_1}

Answer C text: {AnswerCText\_1}

Answer D text: {AnswerDText\_1}

Answer: {CorrectAnswer\_1}

Incorrect answer: {IncorrectAnswer\_1}

Misconception: {MisconceptionText\_1}

...

Example 100:

ConstructName: {ConstructName\_100}

SubjectName: {SubjectName\_100}

Math problem: {QuestionText\_100}

Answer A text: {AnswerAText\_100}

Answer B text: {AnswerBText\_100}

Answer C text: {AnswerCText\_100}

Answer D text: {AnswerDText\_100}

Answer: {CorrectAnswer\_100}

Incorrect answer: {IncorrectAnswer\_100}

Misconception: {MisconceptionText\_100}

"""

Here are examples of the generated questions. The 72B model seems to have high question-generation abilities.

ConstructName: Calculate the circumference of a circle given the radius

SubjectName: Circles

Math problem: If the radius of a circle is \( 7 \) cm, what is the circumference of the circle?

Answer A text: \( 22 \) cm

Answer B text: \( 44 \) cm

Answer C text: \( 14 \) cm

Answer D text: \( 154 \) cm

Answer: B

Incorrect answer: A

Misconception: Thinks circumference is radius x pi

ConstructName: Simplify algebraic fractions by identifying and cancelling common factors

SubjectName: Simplifying Algebraic Fractions

Math problem: Simplify the following algebraic fraction:

\[

\frac{6x^2y}{9xy^2}

\]

Answer A text: \( \frac{2x}{3y} \)

Answer B text: \( \frac{6x}{9y} \)

Answer C text: \( \frac{2xy}{3y^2} \)

Answer D text: \( \frac{6x^2}{9y^2} \)

Answer: A

Incorrect answer: B

Misconception: Cannot identify a common factor when simplifying algebraic fractions

These examples are mathematically valid, and it was crucial to include a large number of examples (100 cases) in the prompt to generate valid questions.

**Misconception Generation**

* Used Qwen2.5-32B-instruct to generate misconceptions. The following prompt was used:

"""You are an expert in mathematics.

Refer to the examples below to identify and describe the misconception that led to the incorrect answer.

Example1

ConstructName: Recognise and use efficient methods for mental multiplication

SubjectName: Mental Multiplication and Division

Math problem: Tom and Katie are discussing ways to calculate\\( 21\\times 12\\) mentally. Tom does\\( 12\\times 7\\) and then multiplies his answer by\\( 3\\); Katie does\\( 21\\times 6\\) and then doubles her answer. Who would get the correct answer?

Incorrect answer: Only Katie

Misconception: Does not correctly apply the distributive property of multiplication

Example2

ConstructName: Multiply a decimal by an integer

SubjectName: Mental Multiplication and Division

Math problem:\\( 9.4\\times 50=\\)

Incorrect answer:\\( 4700\\)

Misconception: When multiplying a decimal by an integer, ignores decimal point and just multiplies the digits

ConstructName:{ConstructName}

SubjectName:{SubjectName}Math problem:{QuestionText}

Incorrect answer:{AnswerText}

Misconception:

"""

**Retriever**

* The retrieval model was evaluated by checking not only MAP@25 but also the top-25 recall
* Retrieval models used for Final Submission
  + Ensemble of the following models
    - Qwen2.5-14B-instruct
      * Misconceptions generated during the Misconception Generation step were also added to the text
      * Loss : MultipleNegativesRankingLoss
      * Fine-tuned with LoRA:
        + LoRA\_rank: 32
        + LoRA\_alpha: 64
      * Trained with 1 batch consisting of 1 positive sample and 47 negative samples for each question-answer pair.
        + Increasing the number of negatives was critical.
        + Negatives used for training were restricted to those associated with positive samples within the training data.
        + Negatives were selected randomly.
      * map 25 : 0.511
      * recall 25 : 0.923
    - Qwen2.5-32B-instruct-GPTQ-Int4
      * Almost same as 14B
      * Trained with 1 batch consisting of 1 positive sample and 255 negative samples for each question-answer pair.
      * 10 epochs (each fold took 10 hours on an A100).
      * Due to the long inference time (50min/fold), only 2 folds were used for the final submission.
      * map 25 : 0.554
      * recall 25 : 0.926
      * map 25 for generated data: ~0.7
      * recall 25 for generated data: ~0.99
* **Inference During Submission**
  + vllm was utilized to speed up the embedding calculation. Since it could not be used directly, some modifications were made to its implementation to adapt it to our use case. This approach enabled efficient inference while maintaining accuracy.
  + The embeddings from the above models were concatenated to generate the final representation.
  + Number of Retrieved Items
    - 25 items from all misconceptions.
    - 15 items from misconceptions not present in the training data:
      * This was prioritized because most misconceptions in the test set were predicted to be absent in the training data.
    - Duplicates between the above were removed.

**Reranker**

* Performed inference using three final models.
* **Model 1 & Model 2**
  + Qwen2.5-32B-Instruct-GPTQ-Int4
    - Trained across 4 folds.
    - Fine-tuned using LoRA.
    - Used a ratio of 1 positive : 9 negatives during training.
    - Limited negatives to those present in the positives.
      * This approach improved CV.
  + **Fold 1 results:**
    - MAP@25: 0.653
    - Evaluated using only misconceptions not present in the training data: MAP@25: 0.601
  + **Model 1:**
    - Ensembled the LoRA components of fold 1 and fold 2.
  + **Model 2:**
    - Ensembled the LoRA components of fold 3 and fold 4.
* **Model 3**
  + Qwen2.5-32B-Instruct-GPTQ-Int4
    - In addition to the competition data, 2500 generated samples were used for training.
    - Trained across 2 folds and ensembled the LoRA components of them.
    - Used a ratio of 1 positive : 19 negatives during training.
    - 2epochs (each fold took 10 hours on 4 x A100).
  + **Fold 1 results:**
    - MAP@25: 0.664
    - Evaluated using only misconceptions not present in the training data: MAP@25: 0.605

The private score was boosted by 0.01 through ensembling a reranker model (Model 3) trained with data generated by the 72B model.

**PostProcess**

* Adjusted the predictions for misconceptions that existed in the train set by reducing their scores:
  + Prediction Score \* 0.40
    - Other coefficients were tested, but 0.40 resulted in the highest public score.